THE PHYSICAL SOCIETY

REPORTS ON PROGRESS IN PHYSICS

VOLUME XXII (1959)

EXECUTIVE EDITOR
A. C. STICKLAND

Published by
THE PHYSICAL SOCIETY
1 Lowther Gardens, Prince Consort Road
London S.W.7
1959

Printed in Great Britain by John Wright and Sons Ltd.
The Stonebridge Press, Bath Road, Bristol 4

CONTENTS

	PAGE
PREFACE	iv
Applications of Quantum Mechanics in Theoretical Chemistry. By G. G. HALL	1
Progress in Vacuum Technology. By J. POLLARD	33
The Propagation of Elastic Waves in Crystals and other Anisotropic Media. By M. J. P. MUSGRAVE	74
Relativity. By H. BONDI	97
Double Beta Decay. By H. PRIMAKOFF and S. P. ROSEN	121
Physical Problems of Thermoelectricity. By A. F. JOFFE and L. S. STIL'BANS	167
Paramagnetic Resonance Data. By J. W. ORTON	204
Experimental Investigations of the Ionospheric E-Layer. By B. J. ROBINSON	241
Radio-frequency Spectroscopy of Excited Atoms. By G. W. SERIES	280
The Theory of a Fermi Liquid. By A. A. ABRIKOSOV and I. M. KHALATNIKOV	329
Invariance in Elementary Particle Physics. By N. KEMMER, J. C. POLKINGHORNE and D. L. PURSEY	368
Nuclear Magnetism in Pure Liquids. By J. G. POWLES	433
Experimental Work with ^{3}He. By V. P. PESHKOV and K. N. ZINOV'EVA	504
The Two-Nucleon Interaction. By R. J. N. PHILLIPS	562

PREFACE

The *Reports on Progress in Physics* is under the general supervision of the Papers Committee of the Physical Society. The Members of the Papers Committee are :

Chairman : Professor R. W. Ditchburn.

Honorary Papers Secretary : Dr. H. H. Hopkins.

Ordinary Members : The President (Mr. J. A. Ratcliffe), Dr. M. Blackman, Professor W. E. Burcham, Dr. E. H. S. Burhop, Dr. A. H. Cooke, Professor S. Devons, Professor C. Domb, Mr. J. Dyson, Professor B. H. Flowers, Professor F. C. Frank, Professor F. Llewellyn Jones, Professor N. Kemmer, Dr. R. Latham, Dr. W. M. Lomer, Dr. K. Mendelssohn, Professor R. E. Peierls, Mr. J. W. Perry, Dr. R. W. B. Stephens, Professor D. H. Wilkinson, Dr. J. M. Ziman.

This volume contains three contributions by Russian scientists. The Papers Committee wishes to record its cordial thanks to the distinguished authors of these papers and to those British scientists who have undertaken translation.

Reports published in the Volume are also being issued as individual articles, at prices depending on their length. Information may be obtained from the offices of the Society.

A. C. STICKLAND,
Secretary–Editor.

APPLICATIONS OF QUANTUM MECHANICS IN THEORETICAL CHEMISTRY

By G. G. HALL†
Quantum Chemistry Group, University of Uppsala, Uppsala, Sweden

CONTENTS

PAGE

§ 1. Introduction ... 2
1.1. Survey of quantum mechanics as used in theoretical chemistry 2
1.2. The mathematical background ... 3
1.3. Different methods of applying quantum mechanics ... 4

§ 2. The calculation method : ' solving the Schrödinger equation ' ... 5
2.1. The molecular orbital method ... 5
2.2. More accurate methods ... 6
2.3. The accurate solution as a series of determinants ... 7
2.4. The choice of atomic orbitals ... 10
2.4.1. Hydrogen-like functions ... 10
2.4.2. Atomic self-consistent field functions ... 10
2.4.3. Laguerre functions ... 11
2.4.4. Slater functions ... 11
2.4.5. Dirichlet functions ... 11
2.4.6. Gaussian functions ... 11
2.5. Survey of results and comparison with experiment ... 11
2.6. Influence of electronic computers ... 13

§ 3. The method of parameters : ' allowing experiment to help theory ' ... 14
3.1. The problem of large molecules ... 14
3.2. Empirical parameters in the molecular orbital equations ... 15
3.3. The molecular orbital theory of conjugated molecules ... 18
3.4. Pitfalls and safeguards ... 22

§ 4. The method of models : ' simplifying the problem ' ... 23
4.1. The use of models ... 23
4.2. Various models ... 24
4.2.1. The free electron model ... 24
4.2.2. The valence bond theory ... 25
4.2.3. Atoms in molecules ... 25
4.2.4. Ligand field theory ... 26
4.2.5. The electrostatic method ... 26
4.2.6. Ionic molecules ... 26

§ 5. The present situation ... 27
5.1. Appreciation of the three approaches ... 27
5.2. Some objectives of current research ... 28
5.3. Wider implications ... 29

Acknowledgments ... 30

References ... 30

† Permanent address : Department of Mathematics, Imperial College, London.

Abstract. Progress in the application of quantum mechanics to study the electronic structure of molecules has, until recently, been disappointing and slow. Many difficulties, both of mathematical method and of numerical analysis, arise when there is a large number of particles in the quantum mechanical system. In this review the different ways of applying quantum mechanics to chemical problems are discussed. The accuracy expected and attained by each method is illustrated, but applications to individual molecules are not reviewed.

The coming of electronic computers has already produced great changes in this subject. In particular, it is now possible to calculate highly accurate wave functions from first principles for a range of molecules. Direct calculations of this kind are not possible for large molecules and, for these, the problem has been to find suitable parameters which remain constant for a series of molecules and can be used to describe their properties. This problem is intimately connected with the problem of interpreting quantum mechanical calculations in a chemically significant way. Quantum mechanics can also be applied to molecules by setting up models which simplify the quantum mechanical equations but retain those features which are most significant for a series of related molecules.

Finally, some of the objectives of current research are briefly reviewed. The wider significance of the subject and its importance for chemistry and the theories of the solid state and the nucleus are also discussed.

§ 1. Introduction

1.1. *Survey of Quantum Mechanics as used in Theoretical Chemistry*

The significance of quantum mechanics for chemistry is that it gives the solution, in principle, to almost every chemical problem. By solving the Schrödinger equation for a molecule the wave functions of all its stationary states can be found and these determine either directly, or indirectly using statistical mechanics, the molecular properties which are studied in experimental chemistry. Quantum mechanics in general may not yet have reached its final form but the quantum mechanics of electrons and nuclei, considered primarily as point charges, is already sufficiently complete and accurate for all chemical purposes. The calculation and prediction of molecular properties involve no new fundamental laws but only the solution of equations. Very unfortunately, however, there is an enormous gap between this solution in principle and the practical calculation of the properties of any specific molecule.

The wave functions of a molecule depend on the positions and spins of all its electrons and nuclei. The Schrödinger equation, which determines them, is usually written by translating the classical expression for the total energy into a sum of differential operators corresponding to the kinetic energies of electrons and nuclei together with the electrostatic potentials of the electron–nuclear attractions and the electron–electron and nuclear–nuclear repulsions.† Because of the large ratio between the nuclear masses and the electron mass it is usually a good approximation to write the molecular wave function as a product of a nuclear wave function and an electronic wave function and so to separate the equation into an electronic equation and an equation for the vibrations and rotation of the nuclei (Born and Oppenheimer 1927, Born and Huang 1954, Liehr

† Relativistic corrections, such as the spin-orbit, change of mass and Dirac δ-function terms, have usually been ignored or treated as small perturbations. They will eventually have to be included in the Schrödinger equation especially for molecules containing atoms heavier than neon.

1957). The electronic equation determines the equilibrium configuration of the molecule and its vibration frequencies and so is the more important chemically. This review will not be concerned with the nuclear motion. (For an analysis of the nuclear motion and further references, see Herzberg 1945, 1950, Sponer and Teller 1941, and Wilson, Decius and Cross 1955.) The equation for the electronic wave function, when atomic units are used, is

$$\mathscr{H}\Psi \equiv \left(-\tfrac{1}{2}\sum_i \nabla_i^2 - \sum_{\alpha i}\frac{Z_\alpha}{r_{\alpha i}} + \tfrac{1}{2}\sum_{i\neq j}\frac{1}{r_{ij}} + \tfrac{1}{2}\sum_{\alpha\neq\beta}\frac{Z_\alpha Z_\beta}{r_{\alpha\beta}} \right)\Psi = E\Psi, \qquad \ldots\ldots(1)$$

where i ranges over the electrons and α over the nuclei. In this equation the relative positions of the nuclei occur as parameters so that both Ψ and the eigenvalue E are functions of these positions. The minimum value of E as a function of the nuclear coordinates determines the equilibrium configuration of the molecule.

There is no general method of solving this electronic Schrödinger equation exactly. Approximate wave functions, however, can always be found using the minimum energy theorem which states that

$$e \equiv \int \Phi^* \mathscr{H} \Phi \, d\tau \Big/ \int \Phi^* \Phi \, d\tau \geqslant E \qquad \ldots\ldots(2)$$

for any approximate wave function Φ. A trial wave function can then be set up containing constants or functions which can be varied and the lowest value of e will give the best approximation to E. It can also be shown (Eckart 1930, Kinoshita 1957) that, as e becomes close to E, the approximate wave function approaches the true eigenfunction.† The choice of trial wave functions is guided in two ways. One way is by the necessity that all the integrals in e should be relatively easy to evaluate and the other that the form of the wave function should accord with what experience, backed by some general considerations, has shown to be desirable. Thus a trial wave function should always be an eigenfunction of the total spin operators S_z and S^2, antisymmetrical in the electronic coordinates, and belong to some irreducible representation of the symmetry group of the nuclei.

The parts of quantum mechanics most relevant to theoretical chemistry are introduced in some detail in the textbooks of Pauling and Wilson (1935) and Eyring, Walter and Kimball (1944). The classic book by Pauling (1940) discusses a wide variety of chemical problems in the light of quantum mechanical ideas but without detailed calculations. The more recent texts by Dewar (1949), Coulson (1952), Pitzer (1953) and Kauzmann (1957) are also written from a chemical point of view but do give some of the more elementary applications in more detail. Unfortunately, there is no textbook yet which gives an adequate account of the whole subject. There are valuable surveys of recent work in the various volumes of the *Annual Review of Physical Chemistry*, and these give many references.

1.2. *The Mathematical Background*

The minimum mathematics necessary to apply quantum mechanics to chemical problems is not extensive or difficult but it is not always familiar to a chemist. On the other hand, there are some problems which demand very highly developed

† In more accurate terms, the overlap integral $\int \Phi^* \Psi \, d\tau$ converges to unity.

mathematical ideas and techniques. (For accounts of some of these, see Kemble (1937) and Titchmarsh (1958).)

The most important branch of mathematics for theoretical chemistry is that dealing with eigenvalues and eigenfunctions both of differential operators and of matrices. A knowledge of some practical methods of solving eigenvalue problems for matrices is essential to most calculations. It is also an advantage to know how these problems may be simplified by the use of group representation theory.

Another important idea is that of the expansion of a function in terms of a complete set of functions. This is a generalization of the idea of expanding a function as the linear combination of sines and cosines which is its Fourier series. Similarly in spaces of many dimensions there are sets of functions which are complete in the sense that a function of many variables, such as a wave function, can be expressed as a linear combination of the functions in the set. An important particular instance is the theory of expansions in spherical harmonics.

The orbital angular momenta are not constants of motion for the electronic Schrödinger equation of a molecule.† Consequently the extensive development of the theory of angular momentum operators, which dominates the theory of atoms and the shell theory of nuclei, is not necessary for molecules. The spin angular momenta are still constants, however, so that a part of the theory is needed in order to construct trial functions which are eigenfunctions of S^2 and S_z.

The best introduction to these mathematical ideas and techniques is the textbook by Margenau and Murphy (1956) though the classic treatise of Courant and Hilbert (1953) is still an essential reference.

1.3. *Different Methods of applying Quantum Mechanics*

If it were possible to solve the Schrödinger equation exactly or to obtain good approximate wave functions in analytical or numerical form, the calculation of molecular properties would be relatively easy and chemical problems could be reduced to numerical ones. Unfortunately even very approximate wave functions are very difficult to find and use so that the straightforward application of quantum mechanics is limited to the smaller molecules.

A more powerful method is to consider series of related molecules and express their properties in terms of a limited set of related integrals. These integrals can then be calculated approximately by a calculation on one molecule or they can be treated as unknowns and estimated by comparison with experiment. This is similar to the approach in the study of atoms which concentrates on understanding selection rules and the various splitting ratios rather than the absolute energies and transition moments.

A third method of applying quantum mechanics employs models. A model isolates from the complicated molecular problem certain significant features and studies these in isolation. When a few features dominate the properties of a series of molecules a suitable model may give great insight into these properties and enable calculations to be made very easily. On the other hand, a model involves a simplification of the real problem and so has to be used with care.

† For a linear molecule the angular momentum around the nuclear axis is a constant.

§2. The Calculation Method: 'Solving the Schrödinger Equation'

2.1. *The Molecular Orbital Method*

In the application of quantum mechanics to atoms the idea of a self-consistent field plays a dominant role (see Hartree 1957). The complicated Schrödinger equation for all the electrons simultaneously is broken up into equations for one electron at a time. Each electron moves in a self-consistent field consisting of the nuclear attraction and the averaged repulsions of the remaining electrons. Mathematically this is an approximation and is equivalent to the use of a trial wave function in the form of a determinant with the atomic orbitals as its elements. The determinantal form is the simplest form which satisfies the Pauli principle.

The basic idea of reducing the Schrödinger equation to orbital equations can also be applied to molecules. By taking a trial wave function in the form

$$\Phi = (2N!)^{-1/2}\,\mathrm{Det}\,\{\psi_1(1)\,\alpha(1);\ \psi_1(2)\,\beta(2);\ \ldots;\ \psi_N(2N)\,\beta(2N)\}, \qquad \ldots\ldots(3)$$

the equations for the molecular orbitals ψ_s can be shown (Lennard-Jones 1949) to be

$$\{-\tfrac{1}{2}\nabla^2 + V(\mathbf{r})\}\,\psi_s(\mathbf{r}) = E_s\psi_s(\mathbf{r}), \qquad \ldots\ldots(4)$$

with the potential $V(\mathbf{r})$ determined from the nuclear attractions and the averaged electronic repulsions, viz.

$$V(1)\,\psi_s(1) = \left\{-\sum_\alpha \frac{Z_\alpha}{r_\alpha} + 2\sum_n \int \frac{\psi_n^*(2)\,\psi_n(2)}{r_{12}}\,dv_2\right\}\psi_s(1) - \sum_n \psi_n(1)\int \frac{\psi_n^*(2)\,\psi_s(2)}{r_{12}}\,dv_2. \qquad \ldots\ldots(5)$$

The corresponding differential equation for an atomic orbital can be solved by numerical integration since V depends only on the radius but here V is much more complicated and numerical integration is too difficult. Instead, the equations can be solved approximately by expressing each molecular orbital as a linear combination of some fixed functions $\omega_l(\mathbf{r})$ so that

$$\psi_s(\mathbf{r}) = \sum_{l=1}^{M} a_{sl}\,\omega_l(\mathbf{r}). \qquad \ldots\ldots(6)$$

If the functions ω_l are drawn from a complete set then, as M is increased, these molecular orbitals should converge to the best molecular orbitals. The coefficients a_{sl} are determined by the equations (Hall 1951, Roothaan 1951),

$$\sum_r^M a_{sr}\left\{H_{lr} + \sum_{jk}^{M}\sum_n^N a_{nj}^*\,a_{nk}[2(lj|G|rk) - (lj|G|kr)]\right\} = E_s\sum_r^M a_{sr}\,S_{lr} \qquad \ldots\ldots(7)$$

where

$$H_{lr} = \int \omega_l^*\left(-\tfrac{1}{2}\nabla^2 - \sum_\alpha \frac{Z_\alpha}{r_\alpha}\right)\omega_r\,dv, \qquad \ldots\ldots(8)$$

$$S_{lr} = \int \omega_l^*\,\omega_r\,dv, \qquad \ldots\ldots(9)$$

$$(lj|G|rk) = \int \omega_l^*(1)\,\omega_j^*(2)\,(1/r_{12})\,\omega_r(1)\,\omega_k(2)\,dv_{12}. \qquad \ldots\ldots(10)$$

These equations are cubic in the a_{sr}, but they can be solved iteratively,† and in practice the use of the symmetry of the molecule brings some simplifications. For a molecule, therefore, the self-consistent field is unimportant and it is the molecular orbitals themselves which are self-consistent.

In principle the functions $\omega_t(\mathbf{r})$ could be chosen from any complete set of functions but, in practice, to ensure rapid convergence they have to be selected carefully. The functions most frequently used are the atomic orbitals of the various atoms in the molecule or some approximation to them. The whole method is then known as the linear combinations of atomic orbitals (LCAO) method. Another choice, which has obvious advantages for diatomic hydrides and similar molecules, is of the atomic orbitals of the heaviest atom. This may be called the single centre method. The factors which determine the choice of atomic orbitals are discussed in §2.4. It is noteworthy, however, that the first rigorous calculation of a molecular orbital (for H_2 by Coulson (1938)) used functions related to the molecule as a whole and not to the component atoms.

2.2. *More Accurate Methods*

Even with the best possible molecular orbitals the molecular orbital wave function is of limited accuracy. Its determinantal form ensures that the Pauli principle is satisfied and consequently the electrons with the same spin are kept apart but some of those with different spins can become close—more close than their Coulomb repulsion permits in reality. This lack of correlation between electrons of different spins is the greatest fault of the molecular orbital method. It is most marked for the two electrons which share the same molecular orbital.

The method of different orbitals for different spins is a generalization of the molecular orbital method which tries to remedy a major part of this defect of correlation without losing the simple physical idea of electrons moving independently in orbitals. Instead of having two electrons with opposite spins in the same spatial orbital they are given different spatial orbitals. These orbitals are not orthogonal to one another and their overlap integrals, unfortunately, make the analysis more complicated. Another mathematical difficulty is that a single determinant of these orbitals is not usually an eigenfunction of the spin operator S^2. This can be overcome using a powerful projection operator technique (Löwdin 1955) which purifies the determinant and yields an appropriate eigenfunction of S^2 as a sum of determinants.

The use of different orbitals for different spins still does not give an exact solution of the Schrödinger equation although it represents a considerable improvement over the molecular orbital method. The attempt has naturally been made to calculate wave functions which converge to the exact solution by methods which are successful for atoms. The most noteworthy of these is that of Hylleraas (1929) whose trial wave function for helium is a series depending explicitly on the interelectronic distance r_{12}. In a similar way, James and Coolidge (1933) have calculated a wave function for H_2 which depends explicitly on r_{12} and has an energy in close agreement with experiment. This dependence on r_{12} speeds the

† It is important that the iterative process used should be second order since the first order process frequently diverges ; see, for example, Hartree (1952), p. 196.

convergence of the wave function but it leads to mathematical and numerical difficulties when attempts are made to generalize the method to systems containing more electrons.

2.3. *The Accurate Solution as a Series of Determinants*

The only general method discovered so far of finding a wave function which converges to the eigenfunction of the Schrödinger equation is by expressing it as a series of determinants. The elements of the determinants are one-electron functions, and if these are chosen from a complete set then the set of all possible determinants that can be constructed using them is complete for any antisymmetrical wave function (see, for example, Kramers 1957). Thus, if a typical determinant is denoted by

$$D_r = (N!)^{-\frac{1}{2}} \operatorname{Det}\{\phi_{r1}(1)\,\alpha(1), \ldots, \phi_{rN}(N)\,\beta(N)\}, \qquad \ldots\ldots(11)$$

where the ϕ_{ri} are some selection of one-electron functions, the most general trial wave function has the form

$$\Psi = \sum_r a_r D_r, \qquad \ldots\ldots(12)$$

where the constants a_r remain to be found by minimizing the energy.

The number of independent constants in this expansion of the wave function can be considerably reduced by using other properties of the wave function. The spin and symmetry properties of the wave function are usually known in advance and so the series can be restricted to those determinants or sums of determinants which have the same properties. This may conveniently be done by using a projection operator $O(S^2)$ to select out a function which is an eigenfunction of S^2 with specified eigenvalue (Löwdin 1955), $O(S_z)$ to select an eigenfunction of S_z, and $O(G)$ to select a function adapted to the symmetry group of the nuclei for a specified irreducible representation (Melvin 1956). The expansion then becomes

$$\Psi = \sum_s a_s O(G)\, O(S^2)\, O(S_z)\, D_s, \qquad \ldots\ldots(13)$$

where now the number of terms in the summation is much less.†

A wave function in this general expanded form is very cumbersome to handle and impossible to interpret in simple physical terms. It has, therefore, been of great importance that quantities known as density matrices can be introduced (Löwdin 1955) which have some immediate physical significance and make calculations easier and more systematic. These density matrices are completely determined by the normalized wave function Ψ. The most important are the two-electron density matrix, defined as

$$\Gamma(1'\,2'\,|\,12) = \tfrac{1}{2}N(N-1)\int \Psi^*(1', 2', 3, \ldots, N)\,\Psi(1, 2, 3, \ldots, N)\, d\tau_3 \ldots d\tau_N, \qquad \ldots\ldots(14)$$

† Boys (1950) has coined the terms ' detor ' for a determinant of orthonormal orbitals such as D_s and ' co-detor ' for a linear combination of determinants coupled to give an eigenfunction of the appropriate operators such as $O(G)\,O(S^2)\,O(S_z)\,D_s$.

in which the numbers represent all the coordinates of the corresponding electron, and the one-electron density matrix

$$\gamma(1'|1) = N\int \Psi^*(1', 2, 3, \ldots, N)\,\Psi(1, 2, 3, \ldots, N)\,d\tau_2 \ldots d\tau_N,$$

$$= \frac{2}{N-1}\int \Gamma(1'2|12)\,d\tau_2. \qquad \ldots\ldots(15)$$

The diagonal elements of these matrices have a simple physical meaning since $\gamma(1|1)$ is the total electron density and $\Gamma(12|12)$ is the correlation density, i.e. it gives the probability of one electron being at the point 1 and another at 2. In terms of these matrices the energy corresponding to Ψ takes the simple form†

$$E = \tfrac{1}{2}\sum_{\alpha\beta}{}' \frac{Z_\alpha Z_\beta}{r_{\alpha\beta}} + \int\left\{\left(-\tfrac{1}{2}\nabla_1^2 - \sum_\alpha \frac{Z_\alpha}{r_{1\alpha}}\right)\gamma(1'|1)\right\}_{1'=1} d\tau_1 + \int\left(\frac{1}{r_{12}}\right)\Gamma(12|12)\,d\tau_1\,d\tau_2. \qquad \ldots\ldots(16)$$

There are two basic types of one-electron function ϕ_r which can be used to set up the determinants D_r. The first is the molecular orbital type consisting of the molecular orbitals as found above and supplemented by other functions orthogonal to them and similarly adapted to the symmetry of the molecule.‡ The advantage of this type is that the molecular symmetry can be satisfied at an early stage in the calculation and that the molecular orbitals are orthonormal even though the fixed functions ω_l used to find them need not be. A wave function built up in this way may be called a converging wave function based on molecular orbitals.§ It has also the advantage that the largest terms in the expansion are, very probably, those which differ from the basic molecular orbital determinant D_0 only in the functions of two electrons (Slater 1929, Brillouin 1933).

The other basic type of function consists of approximate atomic orbitals supplemented by functions also localized around an atom. The advantage of this type is that the determinants can be written down directly without the intermediate step of finding the molecular orbitals. This may be called a converging wave function based on atomic orbitals though it is usually called an electron-pair (or Heitler–London) wave function with excited structures.‖ The non-orthogonality of the usual atomic orbitals is here a major difficulty. It is simplest to begin the calculation by making the orbitals orthogonal, but this has the consequence that a large number of determinants D_r must be included in the expansion before the wave function becomes a reasonable approximation (Slater 1951, McWeeny 1954). Alternatively, the overlap integrals can be retained in the calculation though the formulae are made very much more complicated (Löwdin 1955).

† In the first integral the Laplacian acts only on the unprimed variable. After differentiation the primed variable is equated to the unprimed.

‡ The 'unoccupied' solutions of the molecular orbital equations are often used, but this is not necessary and is even undesirable, if pressed too far, since these solutions are not complete.

§ It is often called a molecular orbital wave function with configuration interaction, particularly when the expansion is terminated after a few terms, but this can be confusing since configuration interaction implies, in atomic spectra, a rather different phenomenon.

‖ If all the possible D_r that can be formed from a fixed number of ω_l are included in the expansions, the two converging wave functions will give exactly the same results despite their differences in form (Slater 1930, Longuet-Higgins 1948).

The process of calculating a converging wave function can be divided into a number of steps. The initial step is to choose the functions ω_t (for details of this see § 2.4). Then the one-electron functions ϕ_r are built up from these, by solving the molecular orbital equations if a molecular orbital basis is chosen or by orthogonalization if the basis is atomic. Next, a number of determinants of type D_r are selected and the projections carried out to give the various terms in the expansion of Ψ. The practical success of the calculation often depends on a good selection of these D_r and some general arguments can be used to guide the choice. Chemical arguments have sometimes been used but these are not reliable.

The second major step is to use this trial wave function to find the density matrices. Since the wave function is a linear expansion in the unknown a_s, these matrices are quadratic forms in the a_s and the coefficients of each product $a_s a_t$ have to be found separately. This reduction from integrals over N electrons to integrals over 1 or 2 electrons is often the longest step in the calculation.

The third step is the evaluation of the one- and two-electron integrals. If all the functions ω_t are expressed in terms of coordinates around one centre, the integration is relatively easy (see, for example, Löwdin 1948, Boys 1950), but if several centres are involved the integration becomes very complicated. A review of integration formulae and tables has been given by Dalgarno (1954) and there are tabulations, by Preuss (1956) and Kotani, Amemiya, Ishiguro and Kimura (1955), of some of the more useful integrals. In general, it is necessary next to transform the integrals over the functions ω_t into integrals over the functions ϕ_r.

The energy is a quadratic form in the a_s and the final step is to calculate its coefficients by multiplying the one- and two-electron integrals by the elements of the density matrices. The minimization of the energy is then a matter of diagonalizing this form. The lowest root of the determinantal equation gives the best approximation to the ground state energy and the corresponding eigenvector gives the values of the a_s for the wave function. The next lowest root gives an approximation to the energy of the first excited state of the molecule whose wave function has the same spin and symmetry characteristics as the ground state (Hylleraas and Undheim 1930, MacDonald 1933).

This long process of calculation yields a wave function in the form of a very general expansion which is difficult to interpret or to discuss in physical terms. The difficulty is due primarily to the fact that the wave function is by no means uniquely expressed. Two wave functions may give exactly the same results and so represent the same solution but have completely different forms and coefficients. For this reason the interpretation is best given in terms of the various density matrices. The first-order density matrix, for example, is a quadratic form in the orbitals ϕ_r

$$\gamma(1'|1) = \sum_{rs} a_{rs} \phi_r^* \alpha(1') \phi_s \alpha(1), \qquad \ldots\ldots(17)$$

where the a_{rs} can now be computed numerically and the sum includes all arrangements of the spin functions. This quadratic form also can be diagonalized and its eigenfunctions are the natural spin orbitals (Löwdin 1955, Löwdin and Shull 1955) $\chi_r(1)$ so that

$$\gamma(1'|1) = \sum_{r} n_r \chi_r^*(1') \chi_r(1). \qquad \ldots\ldots(18)$$

This gives a very simple interpretation to the electron density since it is now the sum of one-electron densities but with occupation numbers n_r which are no longer integers. If the n_r are all different, the natural spin orbitals are uniquely defined and so give a means of comparing and interpreting wave functions. Similar arguments can be used to derive the natural two-electron functions from the second-order density matrix.

2.4. *The Choice of Atomic Orbitals*

The initial step in almost every calculation is the choice of the fixed one-electron functions ω_i. In principle, if these are chosen from a complete set of functions, the choice is irrelevant but, in practice, the rate of convergence depends so heavily on the choice that this has to be considered carefully.

The requirement that the set of functions should be complete has sometimes been thought unnecessary mathematical pedantry, but experience has shown its practical importance. An incomplete set is much less flexible than a complete set so that its use leads to slower convergence and usually to a false limit. On the other hand, not every complete set is suitable. A major requirement is that the scale of the functions should be in the correct relation to the problem. It is important not only that the first few terms should be a good representation of the final function, but also that they should permit the function to be adjusted in the most significant regions. A further requirement is that the integrals which arise should be relatively easy to evaluate. It is this which has led to the frequent choice of the ω_i as atomic orbitals with the spherical harmonics

$$P_l^m(\cos\theta)\,e^{im\phi}, \qquad \ldots\ldots(19)$$

as the angular parts and one of the following types of radial function.

2.4.1. *Hydrogen-like functions.*

The radial functions of a hydrogen-like atom of charge Z have sometimes been used. These are of the form

$$r^l L_{n+l}^{2l+1}(2Zr/n)\,e^{-Zr/n} \qquad \ldots\ldots(20)$$

and have the advantage of being orthogonal. Their disadvantage is that they are not complete unless the continuum functions are also included. Furthermore, successive functions spread over a larger and larger distance and so cease to be significant.

2.4.2. *Atomic self-consistent field functions.*

It has often been considered that the best functions to use would be close to the self-consistent orbitals for the atoms since molecular formation is such a small effect compared with atomic binding. These atomic functions are usually known only as numerical tables so that their use in a molecular calculation is difficult and they have the disadvantage, since they are finite in number, of needing to be supplemented by some other orbitals. Evidence is now growing that this choice is not as good as was hoped. The modification of the orbitals in the molecule may be considerable especially in regions between atoms. Furthermore, the accuracy of the numerical functions is not sufficient for calculations of total energies.

2.4.3. *Laguerre functions.*

The associated Laguerre functions

$$r^l L_{n+l+1}^{2l+2}(2\alpha r)\, e^{-\alpha r} \qquad \ldots\ldots(21)$$

are very similar to the hydrogen-like functions. The fact that the exponent α is held constant is sufficient, however, to make the set complete and to ensure that the functions do not expand too rapidly. They are also orthogonal and many of their integrals can be found systematically using the generating functions.

2.4.4. *Slater functions.*

The most widely used functions are the Slater orbitals

$$r^{n+l} e^{-\alpha r} \quad (n = 0, 1, 2, \ldots), \qquad \ldots\ldots(22)$$

where α may be kept fixed but is usually allowed to vary. This choice has given fairly rapid convergence and been most useful for atoms (see, for example, Coulson and Duncanson 1944). For molecules it is not so satisfactory since the complexity of the molecular integrals increases rapidly with n.

2.4.5. *Dirichlet functions.*

The simple functions

$$r^l \exp(-\alpha_n r), \qquad \ldots\ldots(23)$$

where the $1/\alpha_n$ form a divergent series, such as $\alpha_n = kn$, seem to be free from major disadvantages. Because of their use in Dirichlet series (see, for example, Hardy and Riesz 1915) they may be called Dirichlet functions. They have the advantage that the α_n may be selected at any stage in the calculation to improve the function either for small or large r. In common with the Slater functions these functions are not orthogonal.

2.4.6. *Gaussian functions.*

All the functions above suffer from the disadvantage that the integrals which involve functions on several centres are difficult to evaluate. For this reason the Gaussian functions

$$r^{n+l} \exp(-\alpha r^2) \quad (n = 0, 1, 2, \ldots), \qquad \ldots\ldots(24)$$

where α may be kept fixed or varied, have been of great significance since all their integrals can be easily evaluated (Boys 1950). Unfortunately, the convergence to a good wave function is so slow that the use of the functions is very limited.

A complete set of atomic orbitals around one centre is sufficient to describe any molecular function but, unless the molecule has one dominant atom, the convergence may be slow. It is usual, therefore, to work with overcomplete sets and take orbitals from all the centres. This overcompleteness can lead to very ill-conditioned equations.†

2.5. *Survey of Results and Comparison with Experiment*

Once a wave function and energy have been calculated for a molecule it is relatively easy to calculate some of its physical properties. The simplest are perhaps quantities such as the dipole and quadrupole moments which can be

† For a full discussion of the reason for this and of how it can be avoided, see Löwdin (1956), p. 54.

found by integration over the one-electron density matrix.† Other properties, such as the binding energy, dissociation energies, equilibrium internuclear distances and force constants, depend on the energy and require it to be calculated for several configurations of the nuclei. The binding energy could be calculated by subtracting from the calculated energy the experimental atomic energies, which are often known very accurately, but this would involve an undesirable mixture of theoretical and experimental quantities and would force the whole of the error remaining in the total energy into the binding energy. Instead of this, the practice is to calculate the atomic energies using the same atomic orbitals in the hope that the errors in the total energies will be comparable.‡

Properties such as excitation energies, ionization potentials and electron affinities depend on the energies of several different electronic states and may involve a great deal of labour. The energies are calculated for the same configuration of the nuclei (vertical transitions). Similarly, properties such as the electronic transition moments depend on the wave functions of two electronic states and so are difficult to calculate accurately.§ Finally, there are properties such as the electronic polarizability, the diamagnetic susceptibility and the nuclear screening which can be calculated according to the usual perturbation theory formulation, only if a large number of excited wave functions can be found. These properties can often be calculated more efficiently by reformulating the problem as a variational problem (Hylleraas 1930, Hornig and Hirschfelder 1955, McGarvey 1957) but they are still the most difficult kind of property to calculate.

In table 1 the results of several calculations on the hydrogen molecule are compared with each other and with experiment. The first is a simple molecular orbital wave function (Coulson 1937) in which the orbital is the sum of two hydrogen 1s atomic orbitals with an arbitrary exponent. The second is the best possible molecular orbital wave function (Coulson 1938). The third has different orbitals for the two electrons but each consists merely of two exponential functions one on each atom (Weinbaum 1933, Coulson and Fischer 1949). Finally, there is the James and Coolidge (1933) function as an expansion in r_{12}. It can be seen from this table that the improvements obtained in the agreement with experiment as the function becomes more complicated are appreciable but the largest improvement is in the reliability of the calculated values and this is not just as easy to see. Some calculated properties have been included in the table which are virtually inaccessible to experiment. Others of this type, such as the nuclear quadrupole coupling constant whose value enables the nuclear quadrupole of D to be determined from experiment, were calculated using other wave functions and so are not included.

The hydrogen molecular ion H_2^+ is the only molecule whose properties are known more accurately from calculation than from experiment (see Bates, Ledsham

† According to the Eckart (1930) theorem the approximate wave functions, and hence the density matrices, converge, in the mean, to the true ones. This is not sufficient to ensure the convergence of the expectation value of an unbounded operator such as a moment and some further conditions are required.

‡ The fact that binding energies calculated in this way are generally too small is not due to the minimum energy principle but to the choice of parameters and orbitals better suited to the atom than the molecule.

§ An instructive example showing some of the difficulties has been given by Bates (1951).

and Stewart 1953, and references therein). The most accurate wave functions for H_2, as table 1 shows, give results comparable in accuracy with experiment where the latter is available. For no other molecules are there wave functions yet of this standard of accuracy. The accuracy of most of the published wave functions is comparable with that of the simple molecular orbital function in table 1. For a list of references, see Löwdin and Yoshizumi (1959). The situation, however, is likely to improve soon.

Table 1. Comparison of Calculated Properties of Molecular Hydrogen

	Simple M.O.	Best M.O.	Diff. orb.	JC	Expt.
Total energy (A.U.)	−1·1282[a]	−1·13318[b]	−1·148[cf]	−1·173465[d]	−1·17390$_9$[e]
Binding energy (ev)	3·499[a]	3·635[b]	4·039[c]	4·7344[d]	4·7464[e]
Internuclear distance (Å)	1·383[a]		1·417[c]		1·40[g]
Vibration frequency (cm^{-1})	4580[a]		4750[c]		4360[g]
Quadrupole moment ($Å^2$)				0·345[h]	
Polarizability along axis ($Å^3$)	0·83	0·71[k]		0·906[l]	
Polarizability ⊥ axis ($Å^3$)	0·66[j]	0·62[k]		0·659[l]	
Dielectric constant				1·0002666[l]	1·000267[l]
Magnetic shielding const. (10^{-6})	−0·53[m]		−0·52[m]		−0·56[n]

JC, James–Coolidge; a, Coulson 1937; b, Coulson 1938; c, Weinbaum 1933; d, James and Coolidge 1933; e, Shull, private communication; f, Coulson and Fischer 1949; g, Herzberg 1950; h, James and Coolidge 1938; j, Bell and Long 1950; k, Steensholt 1947; l, Ishiguro, Arai, Mizushima and Kotani 1952; m, Das and Bersohn 1956; n, Ramsey 1950.

The greatest profit in this straightforward method of applying quantum mechanics lies not in its reproducing the properties of familiar molecules, even if these are experimentally inaccessible, but in its predicting the properties of unstable molecules, ions, radicals and collision complexes. Because of their short lifetimes, experimental knowledge of these is very limited and theoretical calculation may be the only method of investigation. Thus, for example, H_3, the collision complex involved in the reaction of H with H_2, is probably known in more detail from calculation (Griffing and Vanderslice 1955, Boys *et al.* 1956) than from experiment. This kind of application is likely to become more important in the future as intermediates in reactions, in flames and in mass spectrometers come to be more studied.

2.6. *Influence of Electronic Computers*

As soon as it became clear that the success of James and Coolidge for H_2 could not be easily repeated for other molecules, interest turned from the calculation of accurate wave functions to the exploitation of less accurate ones. The most important single factor in reviving this interest in the last few years has been the advent of the electronic computer. Calculations have become feasible which earlier would have consumed a lifetime.

It is not only the arithmetic of wave function calculation which has been speeded up but also its algebra. The tedious calculation of the coefficients in a density matrix involves primarily the manipulation of symbols according to definite rules and this has been successfully programmed. The substitution of the numerical values of the integrals into the density matrix to give the energy matrix

elements is primarily an elaborate sorting operation and this too can be programmed. The current position is that there are programmes which carry out the entire calculation of a wave function automatically, given only the operator's choice of atomic orbitals, of determinants D_r to be included in the expansion and of the nuclear configuration. The human element is still vital both in making these choices, even though they are being guided more and more by computing experience and, of course, in writing the programme.

The programming of molecular calculations has involved considerable changes in the mathematical techniques used and will eventually induce many more. Already, for example, it is clear that it is very much simpler to have a routine which calculates the one- and two-electron integrals for the parameters required than to attempt any tabulation of them. The ease with which the computer can do iterative calculations has also given a new emphasis to the use of functions defined by recurrence relations. Finally, the fact that the eigenvalues and eigenvectors of a matrix of order up to about 15 can be found easily and of much higher orders by various extensions, has rendered unnecessary some of the earlier devices for avoiding this problem.

§3. The Method of Parameters: 'Allowing Experiment to help Theory'

3.1. *The Problem of Large Molecules*

The calculation method of applying quantum mechanics to the study of molecules is subject to a number of disappointing limitations. It seems very likely that there will always be a limit to the size and complexity of the molecules which can be studied in this way because of the rapid increase in the amount of numerical labour involved. As the number of electrons in the molecule increases, the number of atomic orbitals needed to give similar accuracy increases in proportion, but the number of integrals is approximately its fourth power. The labour involved in transforming these integrals into contributions to a matrix element of the energy increases as a much higher power and the number of matrix elements is also increasing since the number of determinants is larger. Then again, since rounding-off errors may accumulate and cancellations are more probable, the standard of accuracy also has to be increased. The conclusion is that as the number of electrons increases the computing time will increase, probably exponentially, and the demands on the memory of the computer and on the accuracy of each step will also increase though more gradually. The future will certainly bring considerable improvements in the speed, memory capacity and accuracy of electronic computers and it is to be hoped that the mathematical methods and the programming will also lead to improvements, but nevertheless it seems unlikely that wave functions of reasonable accuracy will ever be calculated for molecules having three or more nuclei (other than H) and more than 25 electrons. For a wave function as accurate as that of James and Coolidge, the limit is probably closer to 15 electrons.† Present results, of course, are very far from these limits.

There is another limitation on the usefulness of the calculation method in theoretical chemistry and this arises from the difficulty of interpreting a complicated

† It may be possible to extend these limits for a molecule which has high symmetry, e.g. C_6H_6, SiF_4.

wave function in a way which is illuminating and suggestive to a chemist. Direct interpretation of the wave function in terms of hybridization, covalent and ionic character, resonance, etc. is possible, to some extent, but these concepts are not mathematically defined and so lead to many difficulties when pushed too far (Coulson 1951). As already mentioned, the form of the wave function is not uniquely defined and an interpretation which ignores this must depend at some point on subjective choices. An interpretation in terms of density matrices and natural spin orbitals is free from these defects but is still too abstract. A large number of chemical problems are formulated in a way which is very difficult to relate to wave functions. These problems arise not in connection with one molecule but in attempts to generalize about the properties of a series of molecules which have some features in common. Thus, for example, the CH internuclear distance in any molecule in which these atoms are nearest neighbours is approximately constant but has small variations which can be correlated with the other atoms around C. It is easy to see, on general quantum-mechanical grounds, that this result is reasonable, but it is difficult to see how it would be illuminated or made more precise by calculating an accurate wave function or an internuclear distance.

The real problem, therefore, is how to use quantum mechanics to derive general concepts and principles which will enable molecular structure and properties to be understood. Chemists have already discovered many such principles and concepts by empirical generalization but if these could be given a quantum-mechanical basis their limitations could be more precisely defined and their development carried out mathematically. This is a more subtle problem than the straightforward application of quantum mechanics and the methods of solving it are much less familiar. It is not only a matter of searching for the common characteristics of the wave functions of a group of related molecules but of thinking of suitable molecular concepts which may perhaps only be defined strictly for certain approximate wave functions but may nevertheless remain valid more generally within appropriate limits of accuracy.

Among the problems which this generalized application of quantum mechanics has to solve, the most important is to find suitable localized entities which can be used as invariant units to build up the properties of large complicated molecules from smaller simpler ones. It is very unlikely that such entities will emerge from a study of the accurate wave function and consequently it will not be easy to define them. If it can be shown that there are such entities then their properties can be investigated in two different ways. Existing calculations can be re-interpreted in terms of these entities and their properties found by the calculation method. Alternatively, it is possible to treat integrals over these entities as parameters and find their values by comparison with experiment. This method of parameters is, of course, familiar in other applications of quantum mechanics but the problem of finding suitable parameters has not usually been as difficult.

3.2. *Empirical Parameters in the Molecular Orbital Equations*

The differential equations (4) define uniquely the best molecular orbitals for a molecule. Even though the molecular orbitals are not found in practice by a direct solution of these equations, they can still be defined in this way. From the

examples quoted in table 1 and a number of similar calculations, it is likely that, in general, the best molecular orbital wave function will give a total energy which is about 99% of the experimental total energy. This accuracy is not high enough to give a reliable binding energy but it is probably sufficient for some other purposes. A major part of the remaining energy is associated with the inner shells around the nuclei and these are not usually concerned in chemical properties. Thus a molecular orbital wave function should give some properties to high accuracy, others to a good accuracy only if care is taken to ensure the cancellation of some errors, while some will always be poorly estimated.

A determinantal wave function does not have uniquely defined orbitals since it is unchanged both in form and in magnitude by a unitary transformation of the orbitals. The differential equations defining the transformed orbitals χ_n have the form

$$\{-\tfrac{1}{2}\nabla^2 + V(\mathbf{r})\}\chi_n = \sum_s e_{sn}\chi_s, \qquad \ldots\ldots(25)$$

$$e_{sn} = \int \chi_s^*\{-\tfrac{1}{2}\nabla^2 + V\}\chi_n \, dv, \qquad \ldots\ldots(26)$$

where V is the same potential as before and the e_{sn} are constants. The uniqueness of the molecular orbitals arises, therefore, from the additional condition that this matrix of constants is diagonal. A different set of orbitals, the equivalent orbitals, can be defined by the additional condition that the unitary transformation should be chosen so that the total self-energy of the orbitals

$$\sum_n \iint \chi_n^2(1)\,(1/r_{12})\,\chi_n^2(2)\,dv_1\,dv_2 \qquad \ldots\ldots(27)$$

should be as large as possible. The molecular orbitals and the equivalent orbitals for a molecule are related by a unitary transformation so they both lead to the same wave function but they satisfy different additional conditions and so may have different physical significance.

Molecular and equivalent orbitals differ considerably in properties. The molecular orbitals, as might be expected of the eigenfunctions of the one-electron Hamiltonians, are the orbitals least changed by any disturbance of the molecule. An important example of such a disturbance is ionization. The removal of one electron makes a major change in the molecular wave function, but it makes little change in the molecular orbitals. The ionization potential is the energy of the molecular orbital $(-E_n)$ and the distribution of positive charge is given by its square (Koopmans 1933, Hall and Lennard-Jones 1950). The molecular orbitals belong to the irreducible representations of the molecular group and in general are spread throughout the molecule.

The equivalent orbitals, on the other hand, are localized. The self-energy (27) is one measure of this localization and when it is maximum the ' exchange ' contribution to the total energy is a minimum and the charge clouds interact according to the classical Coulomb repulsion. In a symmetrical molecule the definition implies that these orbitals are equivalent in the sense that they permute into one another. Thus in CH_4, for example, the four outer equivalent orbitals differ only in their orientation and each has the symmetry of one of the four equivalent

C_{3v} subgroups of the molecular group (Hall 1950). This is the same symmetry as that of the chemical bonds.

This chemical correspondence is a general property of equivalent orbitals. In a saturated molecule there are three types of equivalent orbitals : (i) the inner shell orbitals, which form the atomic core and have little chemical significance, (ii) the lone pair orbitals, which are also concentrated around one nucleus but are in the outer part of the atom and may be concentrated in a definite direction, (iii) the bond orbitals, which are concentrated around two nuclei and have the same symmetry as the bonds in the classical valence structure of the molecule. The localization of these orbitals and their chemical significance have been demonstrated most directly by using calculated molecular orbitals to find the equivalent orbitals.† It can also be shown from the differential equation (25), for, when the ' exchange ' part is small, the potential $V(\mathbf{r})$ is large only in the neighbourhood of the localized orbital itself and the attraction of distant nuclei is screened by the repulsion from the electrons in localized orbitals around them. For this reason, too, the equivalent orbital and its properties are determined very largely by its own nucleus or nuclei and are independent of distant parts of the molecule. In contrast to this, the distinctive feature of a conjugated molecule is that it contains a fourth kind of equivalent orbital which cannot be localized‡ to less than three atoms and so is much more sensitive to the environment.

It is natural, therefore, to use these properties of the equivalent orbitals by defining a chemical bond mathematically as a bond equivalent orbital. From this definition the properties to be expected of a bond can be deduced. It can be shown (Hall 1952) that the total energy can be divided up into the sum§ of atomic energies and bond energies in agreement with the well-known experimental fact. By differentiating the total energy it follows also that the internuclear distances and force constants should be constant. The empirical fact that these quantities are such good constants, not only in one molecule but over a large range of molecules of very different constitution and size, is evidence of how localized the equivalent orbitals must be. The total dipole moment can also be expressed as a sum of localized contributions, but here the lone pair contributions are, in general, more important than the bond contributions (Coulson 1951, Pople 1950).

Unfortunately, the simplest form of the theory does not lead to very many new results when used, according to the method of parameters, to analyse experimental data. The reason is that it predicts that many quantities, such as those mentioned above, are constants and these constants have already been discovered empirically. A more refined form of the theory could be developed to predict the small variations in these constants due to environment or distortion. Molecular ionization potentials are an exception to this general rule since, although the e_{ns} matrix elements between equivalent orbitals should be constants, the eigenvalues E_s of this matrix which give these potentials will depend on the size and constitution of the molecule. By taking the e_{ns} as parameters to be determined from experiment, the variation of ionization potential from molecule to

† Details and diagrams for CO_2 are given by Lennard-Jones (1952), CO by Sahni (1953), H_2O by Duncan and Pople (1953), and Ellison and Shull (1955).

‡ ' Localized ' refers here to the region inside which the electron density is appreciable.

§ This is single summation whereas the original energy expression contains double summations.

molecule can be predicted. This application of the theory has now been tested on a wide range of molecules and gives good agreement with experiment (Hall 1951, 1953, 1954, Franklin 1954).

Even in its most refined form the equivalent orbital theory is based on an approximate wave function and so cannot give a wholly accurate interpretation of molecular structure. It has also certain limitations in its usefulness. One of these is that it treats the bonds as units. This prevents it from finding parameters to describe the formation of bonds and their dissociation energies. In fact, any theory based on a single determinant wave function will not be accurate enough to describe these phenomena since they involve the correlation of the two electrons in the bond and, while a single determinant allows adequately for the correlation of electrons in different bonds, it makes no allowance for the correlation inside the bond (see, for example, Lennard-Jones and Pople 1951). Thus it is necessary to look for a more accurate form of wave function which can be interpreted in terms of entities which are more localized than equivalent orbitals, and can be used to discuss the correlation of electrons inside a bond and to derive the bond parameters in terms of some more basic atomic parameters. Attempts have been made to find such a theory using wave functions built up of localized two-electron functions† (Hurley, Lennard-Jones and Pople 1953, Schmid 1953, Parks and Parr 1958), or of many determinants (Manning 1954) or by interpreting the wave function in terms of a localized region (loge) for each electron (Daudel, Odiot and Brion 1954, 1955). The difficulty with all these theories, from the point of view of the method of parameters, is that, as the wave function becomes more complicated, the mathematical expressions also become more complicated and contain more unknown quantities so that no predictions are possible.

3.3. *The Molecular Orbital Theory of Conjugated Molecules*

The greatest success of the molecular orbital theory has been in its application to conjugated molecules, particularly conjugated hydrocarbons. Several reasons have combined to make this so. The great majority of these molecules are flat and their observed properties are principally determined by the small number of electrons which occupy the orbitals antisymmetrical in the molecular plane. Furthermore they are a favourable series for the method of parameters since there is a large number of molecules involving only carbon atoms.‡ It was fortunate too that for conjugated molecules the chemist's idea of a bond was in difficulties and a more theoretical approach was needed.

The first applications of molecular orbital theory to conjugated hydrocarbons (Hückel 1931, Mulliken 1935, Lennard-Jones 1937, Coulson 1939) involved many approximations. The basic assumption was that the mobile electrons§ in antisymmetric orbitals could be treated separately from the remaining electrons which

† Shull (1959) has proposed that these be called geminals.

‡ The very small dipole moment of the CH bond ($<0{\cdot}04$ Debye experimentally) is here very important since it shows that the environment of a carbon atom attached to a hydrogen is almost the same as if it had been attached to another carbon. The orbitals of carbon atoms on the outside of the molecule are therefore very similar to those in the inside. Nor will there be polarization effects due to the presence of strong local dipoles such as occur, for example, when a halogen replaces the hydrogen.

§ Sometimes called π electrons, although this symbol is only applicable to a linear molecule such as acetylene.

formed an inert core.† The molecular orbitals were expressed as linear combinations of hydrogen-like $2p_z$ atomic orbitals on each carbon atom and the coefficients were found by solving a set of linear eigenvector equations. The total energy was the sum of the eigenvalues of these equations and the total wave function a product of the molecular orbitals. Many properties of the molecules could be expressed in terms of an integral β, which was the matrix element between neighbouring atomic orbitals in the eigenvalue equation. This integral could be treated as a parameter and so found empirically but the trend of the results depended almost entirely on the connectivity of the molecule and not on β. Despite all the rather crude approximations, this theory was surprisingly successful and could be used to predict certain properties to a very high accuracy. On the other hand, it failed badly at a number of points. Two notorious examples were its poor predictions of excited states and the very different values of β necessary to fit different properties. Some of the discrepancies were removed when the overlap integrals between neighbouring atomic orbitals were included in the equations, but the resulting mixture of calculated and empirical integrals made it difficult to state the theory clearly and to determine which expressions should be equated to experiment. It also became apparent that the use of a product wave function was highly unsatisfactory. For a review of the theory and a discussion of most of the difficulties, see Mulliken (1949).

This unsatisfactory situation has now been almost completely remedied. It is clear that the fault in the theory lies in its entirely inadequate initial approximations and its success is due to the fact that some of the results are independent of these approximations. The most serious approximation is the use of the product wave functions instead of an antisymmetrized one. For an atom the difference between a product and a determinantal wave function is often small, but for a molecule the difference is usually large (Hall 1956). The exact specification of the multiplicity and symmetry of excited states is also very important and this makes the form of the wave functions and the energy expressions more complicated.

Once the wave function has been fixed in form the theory can be developed rigorously but it takes two different paths according as the calculation method or the method of parameters is used.

If the molecular orbitals are expressed as a linear combination of some set of fixed functions, then they can be determined from the equations (7). In practice these equations are only needed for the antisymmetrized orbitals so that their form is modified slightly. The integrals over the antisymmetrical atomic orbitals are calculated numerically but integrals involving the core are estimated using approximate orbitals and empirical quantities. The molecular orbitals can then be calculated explicitly and the properties of the molecule derived from them (for C_6H_6, see Goeppart-Mayer and Sklar 1938, Parr, Craig and Ross 1950; C_4H_6, Parr and Mulliken 1951). The results of these calculations are not in very good agreement with experiment. This is usually attributed to the use of a single determinant wave function for the ground state, but this approximation is probably not as serious as those of using core integrals which are not well defined and of using only one hydrogen-like $2p_z$ orbital on each carbon atom.

† For some justification of this assumption, see Parr, Ellison and Lykos (1956).

The equations used in calculating these molecular orbitals are quite different in form from those of the Hückel theory. They are cubic, not linear, in the coefficients and have to be solved by iteration. The eigenvalues give the ionization potentials of the molecule and the sum of these is not the total energy. Thus this method makes no contact with the Hückel method. Nor is it easy to formulate it in a suitable form for the method of parameters since the atomic orbitals are not uniquely defined unless some new conditions are imposed on them.

The standard excited state approximation is an attempt both to give some justification for the successful parts of the Hückel theory and to provide a clearly defined basis for a parametric method (Hall 1952). It introduces the standard excited state as a device to enable localized equivalent orbitals to be found. In this rather artificial state the nuclei are held at a constant fixed distance apart but the mobile electrons are excited until they all have parallel spin. This excitation doubles the number of occupied molecular orbitals and it is not surprising to find that the corresponding equivalent orbitals are well localized, one around each carbon atom. These equivalent orbitals, unlike the atomic orbitals, are orthogonal to each other and their matrix elements e_{ns} are negligible except between nearest neighbours.† Thus, for a conjugated hydrocarbon, this matrix can be set up in terms of two parameters, e for its diagonal elements and f between nearest neighbours. The molecular orbitals ψ_n of the standard excited state can then be determined in terms of the standard equivalent orbitals χ_r by solving the eigenvector equations

$$\sum_s e_{ns} T_{rs} = E_r T_{rn} \qquad \ldots\ldots(28)$$

and putting

$$\psi_n = \sum_l T_{nl} \chi_l. \qquad \ldots\ldots(29)$$

Since the molecular orbitals are the ones least disturbed by excitation these molecular orbitals can be used to form determinantal wave functions for the ground state and the various excited states.‡

The vast majority of conjugated hydrocarbons belong to a class known as alternant hydrocarbons (Coulson and Rushbrooke 1940) since their carbon atoms can be numbered so that no two odd nor two even numbers are nearest neighbours. So long as only nearest neighbour effects are included, the molecular orbital theory of these molecules contains many simplifications. The equations (28) reduce to a simpler form (Potts 1953, Hall 1955) and many simple results are accurate to a high degree since correcting terms vanish. In many respects alternants behave as if they had an additional symmetry operation.

The equations (28) are identical in form with those of the Hückel theory although their interpretation is rather different. This is sufficient to ensure that almost all the successful applications of the earlier theory can be reinterpreted. The one-electron density matrix, for example, can be expressed in the forms

$$\gamma(1'|1) = 2 \sum_{n\ \mathrm{occ}} \psi_n^*(1') \psi_n(1) = 2 \sum_{s,l} \sum_{n\ \mathrm{occ}} T_{ns}^* T_{nl} \chi_s^*(1') \chi_l(1)$$

$$= \sum_{sl} P_{sl} \chi_s^*(1') \chi_l(1), \qquad \ldots\ldots(30)$$

† This has been confirmed by calculating the equivalent orbitals for C_4H_4 (Hall 1954).

‡ It follows from this that the standard excited state gives a realization of the compromise Hamiltonian desired by Mulliken (1949). See also Simpson (1958).

and this matrix P_{st} is exactly the familiar charge and bond order matrix (Coulson 1939, Coulson and Longuet-Higgins 1947). Even the numerical values of these bond orders are the same and they are still related to bond length, though because of the more complicated expression for the total energy this relation takes a different form (Deas 1955). Similarly, the calculation of the large diamagnetic anisotropy of these molecules is unaffected (London 1937, McWeeny 1951). The failure of the Hückel theory to give accurate excited states was primarily due to its assumption that the excitation energies were given by the differences of ' orbital energies'. In fact, the lowest triplet state can be calculated in this way (Hall 1952) but in general it is necessary to specify the wave function accurately and to resolve the degeneracies that arise by allowing for symmetry and alternation (Moffitt 1954, Dewar and Longuet-Higgins 1954). Pariser and Parr (1953) and Pariser (1956) show that the excited states of benzene and the polyacenes can then be predicted very accurately.† The ionization potentials can also be accurately calculated (Pople 1953, Hush and Pople 1955, Hedges and Matsen 1958). As an indication of the agreement with experiment obtained some results for naphthalene and anthracene are given in table 2. For a wide range of conjugated hydrocarbons (see, for example, Coulson and Daudel 1957, Hall 1957) the accuracy of the calculations is of this order but often there are few experimental results available for comparison. For substituted hydrocarbons and heterocyclics the calculations depend on additional parameters and are more difficult but are beginning to be successful (see, for example, Murrell and Longuet-Higgins 1955, McWeeny 1957). The chemical reactivity of a molecule is a more difficult property to calculate since many factors enter into it, but by using simple models of the transition states considerable progress has been made in understanding both the rates of reactions and the points of attack (Wheland 1942, Coulson and Longuet-Higgins 1947, 1948, Longuet-Higgins 1950).

Table 2. Calculated and Experimental Properties of Naphthalene and Anthracene

		Naphthalene			Anthracene	
		calc.	expt.		calc.	expt.
Bond lengths (Å)	1, 2	1·370[a]	1·365[b]	1, 2	1·368[a]	1·366[b]
	2, 7	1·401[a]	1·404[b]	1, 13	1·421[a]	1·433[b]
	1, 9	1·415[a]	1·425[b]	9, 13	1·400[a]	1·399[b]
	9, 10	1·426[a]	1·393[b]	13, 14	1·435[a]	1·436[b]
				2, 3	1·406[a]	1·419[b]
Excitation energies (ev)	$^1B_{3u}^-$	4·018[c]	4·0[c]	$^1B_{3u}^-$	3·715[c]	3·8[c]
	$^1B_{2u}^+$	4·493[c]	4·5[c]	$^1B_{2u}^+$	3·648[c]	3·4[c]
	$^1B_{2u}^-$	5·939[c]	5·7[c]	$^1B_{2u}^+$	5·499[c]	5·0[c]
	$^3B_{2u}^+$	2·180[c]	2·5[c]	$^3B_{2u}^+$	1·660[c]	1·7[c]
Diamagnetic anisotropy ($C_6H_6 = 1$)		2·19[d]	2·10		3·45[d]	3·39
Ionization potential (ev)		8·28[e]	8·12[f]		7·38[e]	

a, Deas 1955 ; b, Ahmed and Cruickshank 1952, Cruickshank 1956 ; c, Pariser 1956 ; d, McWeeny 1952 ; e, Pople 1953 ; f, Watanabe 1957.

† The *ad hoc* assumptions on which these authors base their work can be deduced, with one exception to be mentioned, from the standard excited state approximation. In particular, by considering approximate analytic forms for the standard equivalent orbitals it can be shown that the two-electron integrals involving these orbitals are very small unless they are Coulombic.

The difference between the standard excited state theory and the Hückel theory becomes more apparent when refinements to the theory are considered. For a theory based on atomic orbitals the inclusion of overlap integrals was obviously necessary. For a theory based on localized orthonormal orbitals this is unnecessary and refinement consists in improving the wave function for the ground state. This can be done either by expressing it as a sum of determinants using the standard molecular orbitals as the basis (Pariser 1956), or by expressing the ground state molecular orbitals in terms of the standard equivalent orbitals and solving the cubic equations (7) to find the best coefficients (Pople 1953).

3.4. *Pitfalls and Safeguards*

Experience has now shown clearly that there are many pitfalls for the incautious in this way of applying quantum mechanics to molecules by the method of parameters. Some people have even despaired of the possibility of using the method to give any reliable results. This scepticism is not really justified, however, for as the dangers become better understood it is possible to guard against them and to test the reliability of the method by independent means.

One particularly tempting pitfall is the over-hasty identification of a calculated expression with an experimental value. Equations of this kind between quantities which arise from quite different sources require considerable care. It can often happen that the conditions of the experiment are rather different from those envisaged or unconsciously assumed in the calculations. It can also happen that the phenomenon is governed by several factors but that the calculation, through neglect of corrections thought to be small, has ignored some of them. The most obvious illustration of this has been in the calculation of resonance energies of conjugated hydrocarbons. These were among the first properties to be calculated by the molecular orbital and valence bond methods (Hückel 1931) and their excellent agreement with experimental resonance energies gave many chemists their first appreciation of the power and usefulness of quantum mechanics in theoretical chemistry. Unfortunately, this agreement is false because the calculations implied that internuclear distances were the same in the hypothetical Kekulé structure as in the actual molecule. This difference is not negligible and recent empirical estimates of vertical resonance energies show them to be very much larger than the older adiabatic ones. The reason for the good agreement previously found is still not known.

A second kind of pitfall is the use of parameters which are not uniquely defined. It is not sufficient that the parameter be uniquely defined mathematically. It must also be uniquely defined in terms of the observed quantities. Thus, for example, the overlap integral between two atomic orbitals on different atoms can be defined mathematically but not in terms of observed quantities since the orbitals are modified by the approach of the atoms.

Another pitfall is in the continual use of approximations to avoid all the mathematical difficulties in a theory. This is especially tempting when the approximations are in accord with intuitive ideas and the more accurate treatments are not. Individual approximations may be easy to justify but their cumulative effect may be to give quite misleading formulae. The formulae of perturbation theory, for example, may be accurate when interpreted numerically but, when

reduced to the first two terms, may be qualitatively wrong in a method of parameters.

The most notorious pitfall has been the use of an over-simplified theory. It has happened very often that a theory has been developed approximately and is in good agreement with experiment, but when the approximations are refined this agreement is lost. There can be a number of reasons for this behaviour, but one of the most probable is that there is another theory which does not involve these approximations but yet leads to final expressions of the same mathematical form as in the first theory. Several examples of this are mentioned later.

It is possible to guard against these pitfalls in several ways. An obvious way is the rigorous development of the theory. This includes the necessity of stating clearly the assumptions and approximations involved, of seeking to justify them by considering, at least in some examples, the effect of refining them and of verifying that the parameters are uniquely defined and their empirical values of the correct order of magnitude.

A second safeguard is to have an overlap between the method of calculation and the method of parameters. If a wave function of the same form as used in the method of parameters can be calculated accurately it can be used to obtain calculated values of the parameters. These can then be compared with their empirical values and the effect of any approximations assessed quantitatively. The molecules chosen for this purpose will probably be simple so that the overlap also provides a method of extrapolating calculations on small molecules to give results for large ones.

§ 4. The Method of Models: 'Simplifying the Problem'

4.1. *The Use of Models*

The third method of applying quantum mechanics to the problems of molecules has been to use models. In most branches of applied mathematics the use of models to idealize problems and make their solution much easier is fully established. Corresponding developments in molecular quantum mechanics have been slow and difficult, principally because molecules have distinctive nuclei and shapes which are not easy to idealize. Nevertheless there are now a number of models which have had considerable success. Most of these apply only to a limited range of molecules which share some fundamental feature. This key feature is then used to suggest a simplified treatment of the problems.

Because it involves some simplification of the real problem the use of a model is subject to a number of obvious limitations. It cannot be expected to give good results for all molecules and all properties irrespective of whether the simplification is still valid or not. It is clear too that only a limited refinement of a model is worth while. After that it becomes simpler to return to the original Schrödinger equation. There is also the very great danger that a model will prove to be an over-simplified theory so that good agreement with experiment, however great its practical value, is not sufficient by itself to justify the simplifications of the model.

Inevitably the method of models and the method of parameters have much in common and it is sometimes difficult to classify a theory as following the one or

the other. The basic distinction adopted here is between a theory concerned with the Schrödinger equation for the molecule and with a definite wave function, even though it may be very approximate, and one concerned with a different Schrödinger equation, even though it may be solved accurately. This distinction is vital, for the first theory can have its accuracy tested by the calculation method whereas for the second it is difficult or even impossible to do so.

4.2 *Various Models*

4.2.1. *The free electron model.*

It has long been recognized that the mobile electrons of conjugated molecules, which give them their characteristic properties, behave in many respects like the electrons in a metal. It was natural, therefore, to copy the Sommerfeld model of a metal and to treat the mobile electrons as moving freely throughout the molecule. The earliest applications of this model were to cyclic molecules such as benzene and to linear molecules such as the polyene series and these were treated as one-dimensional boxes with lengths determined by the molecules. It has been very successful in explaining such features as the diamagnetic susceptibility (Pauling 1936), the high polarizability (Coulson and Davies 1952), the main features of the electronic spectra (Simpson 1948, 1949, Bayliss 1948, Kuhn 1948) and the nuclear magnetic resonance (Bernstein, Schneider and Pople 1956).

This free electron model has recently been extended to include the whole range of conjugated molecules (Griffith 1953, Ruedenberg 1953, Kuhn 1954, 1956). The electrons are considered as moving freely through the lines which form the framework of the molecule. At a point where the lines meet, their wave functions have to satisfy a number of boundary conditions. Then, by introducing an additional potential in the form of a δ function at the appropriate nucleus, the method can be extended to heterocyclics (Kuhn 1951). The method has given good agreement with experiment for a surprisingly large range of properties.

The success of this model (or of the Sommerfeld box model) certainly cannot be attributed to the accuracy of its assumptions since these include the omission of the electronic repulsions, the restriction to one dimension (contradicting the uncertainty principle) and the use of a constant potential (contradicting the virial theorem). It has now been shown, however, that the model has a very close relation to the molecular orbital method (Griffith 1953, Ruedenberg 1953). If the amplitudes of each free electron orbital at the nuclei are chosen as variables the equations determining them are almost identical with the equations (28) determining the molecular orbitals. It is no accident, therefore, that this method is almost as successful as the molecular orbital method. It follows also that its unrealistic assumptions are irrelevant to its success and merely show that a model should not be expected to be realistic at every point.

A different kind of extension of the free electron model has been given by Platt (1954). In this the electrons move freely in a two-dimensional box whose size and shape are determined by the molecule. This model leads to an electron density whose maxima very surprisingly appear to coincide with the positions of the nuclei.

4.2.2. *The valence bond theory.*

In classical valence theory the properties of a conjugated molecule were interpreted as a resonance between several structures which were alternative ways of arranging the double bonds so as to satisfy all the valencies of the atoms. It was very natural, therefore, that chemists should become very interested when the Heitler–London theory was extended to conjugated molecules and gave a wave function which could be interpreted in terms of the superposition of alternative structures. The rules for calculating with these wave functions could be reduced to a simple form and many properties were studied (for details and further references see Eyring, Walter and Kimball 1944). The theory also served as an effective framework for many qualitative discussions of more complicated molecules in terms of resonance (for details see Pauling 1940).

Unfortunately, in the course of the mathematical development it was assumed that all the atomic orbitals were orthogonal although it was known that, in practice, the overlapping of atomic orbitals on neighbouring atoms was very important. Nevertheless, while the theory was used according to the method of parameters, it appeared to be very successful.† It was only as some complete calculations were carried out (cf. McWeeny 1954) that it was realized that the theory was not justified mathematically without considerable modification.

In its present form, therefore, the valence bond method must be considered as a model of a conjugated molecule which emphasizes the delocalization of the double bonds. The reason for its success is still not wholly explained but it has been suggested (Dewar and Longuet-Higgins 1952) that there is a strong mathematical link between the method of calculation using structures and an approximate solution of the molecular orbital equations. An alternative suggestion (McWeeny 1954) is that the wave function itself is sufficiently accurate but that the overlap integrals need to be included in the energy calculations.

4.2.3. *Atoms in molecules.*

The method of atoms in molecules is an attempt to consider a molecule as a collection of atoms which interact (Moffitt 1951). It divides the Schrödinger equation up into the sum of atomic terms and interaction terms, and uses a wave function based on the occupied atomic orbitals of the atoms and their positive and negative ions.‡ The energy contribution from the atomic terms is evaluated empirically using spectroscopic data, while the interaction part is calculated using approximate forms for the atomic orbitals.

The strongest feature of this method has been its realization that much of the gap between current calculated energies and experiment could be traced to the integrals involving a single atom and that atomic spectroscopy provided a wealth of accurate data about atoms. On the other hand, it is not yet clear that atoms are so little affected by bonding that the atomic data are really relevant.

This method is still in course of development (Hurley 1956, Arai 1957, Preuss 1957) and the mathematical difficulties which arise have not yet been solved. For this reason it is difficult to assess the method and decide whether it involves a

† For molecules larger than anthracene the problem of how to include a sufficient number of structures becomes acute (Pullman 1947).

‡ This wave function is of the Heitler–London type but has both covalent and ionic structures.

model or whether it can be uniquely related both to an accurate calculation and to the experimental data used.

4.2.4. *Ligand field theory.*

Recently the study of complexes in which an atom or ion is loosely bound to a number of surrounding molecules has become of great interest. This binding has been very successfully treated by a model known as the ligand field (or the crystal field). The basic idea is that the influence of the molecules on the atom can be represented by an electrostatic potential. In practice this potential is expanded in spherical harmonics around the atom and only the first few terms in the expansion retained.

The model consists, therefore, of an atom in an external field and this can be solved by the usual atomic methods. The additional potential is readily incorporated into the orbital description but the classification according to angular momenta has to be replaced by a group theory classification based on the symmetry of the external field. In an octahedral field, for example, the five d orbitals split into triply and doubly degenerate orbitals. This splitting gives rise to new phenomena when the orbitals are built up into wave functions for the atom and the energies found. For full references and further details see the recent reviews by Moffitt and Ballhausen (1956) and Jørgensen (1957).

4.2.5. *The electrostatic method.*

The force $\mathbf{F}_\alpha$ experienced by a nucleus α in a molecule due to the other nuclei and the electrons is given, according to the Hellmann–Feynman theorem (Feynman 1939), by the formula

$$\mathbf{F}_\alpha = \sum_{\beta \neq \alpha} Z_\beta \frac{\mathbf{r}_{\beta\alpha}}{r_{\beta\alpha}^3} - \int \gamma(1) \frac{\mathbf{r}_{\alpha 1}}{r_{\alpha 1}^3} \, dv, \qquad \ldots\ldots(31)$$

where γ is the electron density

$$\gamma(1) = \gamma(1|1). \qquad \ldots\ldots(32)$$

This is exactly the classical formula if γ is interpreted as the density of a continuous charge distribution.

This formula (31) has been used to calculate internuclear distances and force constants for various hydrides by using a simple model to estimate the electron density. If the hydrogen nuclei are removed, the problem of the electronic structure reduces to that of the central atom with additional electrons. The model consists in using the electron density from this atom in the formula (31) (Platt 1950, Longuet-Higgins and Brown 1955, Hurley 1954).

This is a very simple theory and yet its results have been very good. Its success depends, at least in part, on correlation errors being the same in the negative ion as in the molecule.

4.2.6. *Ionic molecules.*

Just as ionic crystals can be considered as a collection of ions held together primarily by electrostatic forces, so there is a number of ionic molecules which can be considered in the same way. Their equivalent orbitals can be localized around single centres and correspond to those of the free ions. The structure of

these molecules can be investigated very simply, if the effect of the overlap of the ionic wave functions is represented by a central force. The total energy consists then of the electrostatic energy, the overlap repulsion and the polarization energy and, by minimizing this, the binding energy, internuclear distance, force constant and dipole moment can be predicted and are in good agreement with experiment (Debye 1928, Rittner 1951).

§5. The Present Situation

5.1: *Appreciation of the Three Approaches*

It has not always been recognized that there can be several different methods of applying quantum mechanics to molecular problems and that these have different characteristics. They may often involve different mathematical techniques, but it is more fundamental that they differ in their aims. This gives them different criteria of importance and different standards of success. It may also involve them in different limitations though these are usually due to more obvious mathematical or chemical reasons.

The aim of the calculation method to calculate a good energy and a good wave function is easy to understand and appreciate and has a final quality which is difficult to find in other methods. Its logical status has been much improved by the realization that the calculation can now be carried out entirely in mathematical terms. This has removed the difficulties which kept arising in earlier work when the chemical interpretation was pushed further than is legitimate and used to justify approximations. The mathematical difficulties of the method are still enormous, however, even with the advantage of the revolution in computing brought by the electronic computer. As the structure of the equations becomes better understood, the efficiency of the numerical methods will be increased and some of the technical limitations removed, but many complicated mathematical problems are involved.

The method of parameters has the aim of using the accurate equations as extrapolation formulae to correlate and predict molecular properties accurately. This has often been misunderstood and it has been treated as a semi-empirical method in which mathematical rigour was secondary and agreement with physical and chemical intuition primary. Its agreement with experiment is often surprisingly good but its agreement with the results of pure calculations, when these can be carried through, are just as significant. Its other aim is to provide concepts, less general than that of the wave function itself but more readily interpreted chemically, whereby the variations of molecular properties could be understood. Thus the ideas of orbitals, mobile electrons and alternant hydrocarbons are its contribution to chemical theory. Although the method does give rise to some mathematical problems a greater limitation to its application has been the fact that it requires a number of very accurate experimental values to determine the parameters before any new property or series of molecules can be included in its predictions.

The use of models makes its strongest appeal to chemists and to those who enjoy seeing a wide measure of agreement with experiment obtained with the minimum of effort. Its aim has been to concentrate on some feature which

dominates the properties of a series of molecules. Since it is a simplification of the real equations, the calculations are often relatively simple and, if parameters are introduced, they are few in number and so easily estimated. This has meant that models have become working tools for experimentalists enabling them to interpret results, to estimate the influence of secondary effects and to suggest new investigations. On the other hand, it is never possible to accept a model merely because it leads to good results and without understanding its relation to the true equations. This fuller understanding of a successful model is also a method of investigating the Schrödinger equation itself. The greatest difficulty of the method is in estimating the limits to the validity of a model. If a model is relatively realistic its validity may depend on cancellations of errors and these are difficult to study. If the model contains very unrealistic features it will probably involve an over-simplified theory and its limitations will be bound up with the nature of its correspondence with the more accurate theory.

5.2. *Some Objectives of Current Research*

In a subject which has attracted mathematicians, physicists and chemists and which is viewed at times as a branch of quantum mechanics and at other times as physical chemistry, sometimes as theoretical spectroscopy and sometimes as theoretical chemistry, it would be surprising if there was much agreement about the course of current research. In practice there is often some duplication of work but this is because progress is so slow and easily calculated applications so scarce, and not because of any concentrated drive towards one agreed goal as found in some other subjects.

In current work the calculation method is most in favour because of the many new possibilities opened up by the advent of electronic computers. Wave function calculations of much higher accuracy than before for the simpler diatomic molecules are now being carried out. The calculation of the basic integrals when the atomic orbitals are around three or four different centres is still being studied. The development of the method of different orbitals for different spins is also being studied and so are the possibilities of introducing the interelectronic distances explicitly into the wave function, as has been done so successfully for the two-electron systems.

The use of atoms as a proving ground for methods of calculation has now become firmly established and is leading to a new interest in atomic problems. The calculation of atomic scattering factors (Hurst and Matsen 1958) and electron affinities, for example, have attracted attention and also the cross sections of atoms for various processes (for a review, see Branscomb 1957).

A difficult problem which lies ahead is the re-interpretation of the treatment of electronic correlation given by Bohm and Pines (1951, 1953) and Brueckner (1954, 1955, Brueckner *et al.* 1954) in a way which can be related to molecular calculations.

The current interests in the method of parameters are in the extension of the methods to heterocyclic molecules and substituted hydrocarbons and in applications to properties which have become of importance through the development of new experimental techniques such as nuclear magnetic resonance (Pople 1958) and spin resonance (McConnell 1956).

Recent work on models has been principally concerned with the atoms in molecules method. It is clear that the method is capable of giving good results and this has given a strong motive for finding how to overcome its formal difficulties.

5.3. *Wider Implications*

The profound contributions which quantum mechanics has made to chemistry go far beyond the detailed calculations and theories already mentioned. Of major importance has been its confirmation of the description of matter in terms of electrons and nuclei. The interpretation of experimental results in quantum mechanical terms is also an important contribution. Even if there were no calculated molecular wave functions or energy levels, it would still be important to know that molecular spectra should be interpreted in terms of energy levels and intensities in terms of wave functions. Similarly, in the statistical mechanics of molecules, the formulation in terms of energy levels and wave functions is vital to the success of the theory.

Even the more detailed calculations, however, have wider implications. There is a close parallel, for example, between the methods used in these calculations and those used in the theory of the solid state. This is to be expected since a solid can be considered as the limiting case of a large molecule and, while the crystal symmetry groups play a bigger rôle in the theory, it is not essentially different from that played by the molecular symmetry groups. It has been a major advance in recent years that these two branches of applied quantum mechanics should be recognized as basically one and used to assist one another. In particular, detailed calculations on a molecule can be used to investigate the accuracy of a solid state method and to assess the effect of various approximations.

Another less expected implication of the calculations has been that even the biological properties of molecules can be investigated. The problem of finding the physical properties common to the group of carcinogenic hydrocarbons has been pursued most energetically by the Pullmans (1955) and it is even possible to predict new carcinogens and to suggest some features of the reactions involved.

More generally, the methodology of the subject has many applications in other subjects which are faced with the problems of relating very general, but very abstract, basic laws to experimental results and generalizations. Of wide significance is the conclusion reached that it may be a necessity of thought to find intermediate concepts, which are derived from the basic laws but also involve approximations, and which can be applied in circumstances where the direct application of the basic laws would be impossibly complicated. The experience gained in relating these concepts to the basic laws, to simplified models, and to experimental generalizations should also be valuable. One application of these ideas may be in the theory of the atomic nucleus. It seems clear already that the theory of nuclear forces will be too complicated to be applied directly to calculate the properties of any but the simplest nuclei. On the other hand, there are models which reproduce many important nuclear properties so that the possibilities of finding good intermediate concepts are high.

ACKNOWLEDGMENTS

The author would like to thank the Aeronautical Research Laboratory, Wright Air Development Center of the Air Research and Development Command, USAF, King Gustaf VI Adolf's 70-Years Fund for Swedish Culture, Knut and Alice Wallenberg's Foundation, and the Swedish Natural Science Research Council for their generous support during the preparation of this review.

Thanks are also due to Dr. H. D. Deas, Docent P.-O. Löwdin, Professor F. A. Mátsen and Professor H. Shull, whose careful reading of the manuscript and valuable comments have led to many improvements.

REFERENCES

AHMED, F. R., and CRUICKSHANK, D. W. J., 1952, *Acta Cryst., Camb.*, **5**, 852.
ARAI, T., 1957, *J. Chem. Phys.*, **26**, 435, 451.
BATES, D. R., 1951, *J. Chem. Phys.*, **19**, 1122.
BATES, D. R., LEDSHAM, K., and STEWART, A. L., 1953, *Phil. Trans. Roy. Soc.* A, **246**, 215.
BAYLISS, N. S., 1948, *J. Chem. Phys.*, **16**, 287.
BELL, R. P., and LONG, D. A., 1950, *Proc. Roy. Soc.* A, **203**, 364.
BERNSTEIN, H. J., SCHNEIDER, W. G., and POPLE, J. A., 1956, *Proc. Roy. Soc.* A, **236**, 515.
BOHM, D., and PINES, D., 1951, *Phys. Rev.*, **82**, 625 ; 1953, *Ibid.*, **92**, 609.
BORN, M., and HUANG, K., 1954, *The Dynamics of Crystal Lattices* (Oxford : University Press).
BORN, M., and OPPENHEIMER, J. R., 1927, *Ann. Phys., Lpz.*, **84**, 457.
BOYS, S. F., 1950, *Proc. Roy. Soc.* A, **200**, 542, **201**, 125.
BOYS, S. F., COOK, G. B., REEVES, C. M., and SHAVITT, I., 1956, *Nature, Lond.*, **178**, 1207.
BRANSCOMB, L. M., 1957, *Advances in Electronics and Electron Physics*, **9**, 43.
BRILLOUIN, L., 1933, *La méthode du champ self-consistent* (Act. Scient. et Indust. No. 71).
BRUECKNER, K. A., 1954, *Phys. Rev.*, **96**, 508 ; 1955, *Ibid.*, **100**, 36.
BRUECKNER, K. A., LEVINSON, C. A., and MAHMOUD, H., 1954, *Phys. Rev.*, **95**, 217.
COULSON, C. A., 1937, *Trans. Faraday Soc.*, **33**, 1479 ; 1938, *Proc. Camb. Phil. Soc.*, **34**, 204 ; 1939, *Proc. Roy. Soc.* A, **169**, 413 ; 1951, *Ibid.*, **207**, 63 ; 1952, *Valence* (Oxford : University Press).
COULSON, C. A., and DAUDEL, R. (Editors), 1957, *Dictionary of Values of Molecular Constants* (*Wave Mechanical Methods*) (published privately).
COULSON, C. A., and DAVIES, P. L., 1952, *Trans. Faraday Soc.*, **48**, 777.
COULSON, C. A., and DUNCANSON, W. E., 1944, *Proc. Roy. Soc. Edinb.*, **62**, 37.
COULSON, C. A., and FISCHER, I., 1949, *Phil. Mag.*, **40**, 386.
COULSON, C. A., and LONGUET-HIGGINS, H. C., 1947, *Proc. Roy. Soc.* A, **191**, 39 ; 1948, *Ibid.*, **192**, 16.
COULSON, C. A., and RUSHBROOKE, G. S., 1940, *Proc. Camb. Phil. Soc.*, **36**, 113.
COURANT, R., and HILBERT, D., 1953, *Methods of Mathematical Physics*, Vol. I (London : Interscience).
CRUICKSHANK, D. W. J., 1956, *Acta Cryst., Camb.*, **9**, 915.
DALGARNO, A., 1954, *Math. Tab. Aids Comp.*, **8**, 203.
DAS, T. P., and BERSOHN, R., 1956, *Phys. Rev.*, **104**, 849.
DAUDEL, R., ODIOT, S., and BRION, H., 1954, *J. chim. Phys.*, **51**, 74, 358, 553 ; 1955, *J. Chem. Phys.*, **23**, 2080.
DEAS, H. D., 1955, *Phil. Mag.*, [7] **46**, 670.
DEBYE, P., 1928, *Polar Molecules* (New York : Dover).
DEWAR, M. J. S., 1949, *The Electronic Theory of Organic Chemistry* (Oxford : University Press).
DEWAR, M. J. S., and LONGUET-HIGGINS, H. C., 1952, *Proc. Roy. Soc.* A, **214**, 482 ; 1954, *Proc. Phys. Soc.* A, **67**, 795.
DUNCAN, A. B. F., and POPLE, J. A., 1953, *Trans. Faraday Soc.*, **49**, 217.
ECKART, C., 1930, *Phys. Rev.*, **36**, 878.

ELLISON, F. O., and SHULL, H., 1955, *J. Chem. Phys.*, **23**, 2348.
EYRING, H., WALTER, J., and KIMBALL, G. E., 1944, *Quantum Chemistry* (New York : Wiley and Sons).
FEYNMAN, R. P., 1939, *Phys. Rev.*, **56**, 340.
FRANKLIN, J. L., 1954, *J. Chem. Phys.*, **22**, 1304.
GOEPPART-MAYER, M., and SKLAR, A. L., 1938, *J. Chem. Phys.*, **6**, 645.
GRIFFING, V., and VANDERSLICE, J. T., 1955, *J. Chem. Phys.*, **23**, 1039.
GRIFFITH, J. S., 1953, *Trans. Faraday Soc.*, **49**, 345, *Proc. Camb. Phil. Soc.*, **49**, 650.
HALL, G. G., 1950, *Proc. Roy. Soc.* A, **202**, 336 ; 1951, *Ibid.*, **205**, 541 ; 1952, *Ibid.*, **213**, 102, 113 ; 1953, *Trans. Faraday Soc.*, **49**, 113 ; 1954, *Ibid.*, **50**, 319, 773 ; 1955, *Proc. Roy. Soc.* A, **229**, 251 ; 1956, *Proc. Phys. Soc.* A, **69**, 1124 ; 1957, *Trans. Faraday Soc.*, **53**, 573.
HALL, G. G., and LENNARD-JONES, J. E., 1950, *Proc. Roy. Soc.* A, **202**, 155.
HARDY, G. H., and RIESZ, M., 1915, *The General Theory of Dirichlet's Series* (Cambridge Tracts No. 18).
HARTREE, D. R., 1952, *Numerical Analysis* (Oxford : University Press) ; 1957, *The Calculation of Atomic Structure* (New York : Wiley and Sons).
HEDGES, R. M., and MATSEN, F. A., 1958, *J. Chem. Phys.*, **28**, 950.
HERZBERG, G., 1945, *Infra-red and Raman Spectra of Polyatomic Molecules* (Princeton, N.J. : van Nostrand) ; 1950, *Molecular Spectra and Molecular Structure: 1. Diatomic Molecules* (Princeton, N.J. : van Nostrand).
HORNIG, J. F., and HIRSCHFELDER, J. O., 1955, *J. Chem. Phys.*, **23**, 474.
HÜCKEL, E., 1931, *Z. Phys.*, **70**, 204, **72**, 310.
HURLEY, A. C., 1954, *Proc. Roy. Soc.* A, **226**, 170 ; 1956, *Proc. Phys. Soc.* A, **69**, 49.
HURLEY, A. C., LENNARD-JONES, J. E., and POPLE, J. A., 1953, *Proc. Roy. Soc.* A, **220**, 446.
HURST, R. P., MILLER, J., and MATSEN, F. A., 1958, *Acta Cryst., Camb.*, **11**, 320.
HUSH, N. S., and POPLE, J. A., 1955, *Trans. Faraday Soc.*, **51**, 600.
HYLLERAAS, E. A., 1929, *Z. Phys.*, **54**, 347 ; 1930, *Ibid.*, **65**, 209.
HYLLERAAS, E. A., and UNDHEIM, B., 1930, *Z. Phys.*, **65**, 759.
ISHIGURO, E., ARAI, T., MIZUSHIMA, M., and KOTANI, M., 1952, *Proc. Phys. Soc.* A, **65**, 178.
JAMES, H. M., and COOLIDGE, A. S., 1933, *J. Chem. Phys.*, **1**, 825 ; 1938, *Astrophys. J.*, **87**, 447.
JØRGENSEN, C. K., 1957, *Energy Levels of Complexes and Gaseous Ions* (thesis) (Copenhagen : J. Gjellerup).
KAUZMANN, W., 1957, *Quantum Chemistry* (New York : Academic Press).
KEMBLE, E. C., 1937, *Fundamental Principles of Quantum Mechanics* (New York : McGraw-Hill).
KINOSHITA, T., 1957, *Phys. Rev.*, **105**, 1490.
KOOPMANS, T., 1933, *Physica*, **1**, 104.
KOTANI, M., AMEMIYA, A., ISHIGURO, E., and KIMURA, T., 1955, *Table of Molecular Integrals* (Tokyo : Maruzen).
KRAMERS, H. A., 1957, *Quantum Mechanics* (Amsterdam : North Holland Publ. Co.).
KUHN, H., 1948, *J. Chem. Phys.*, **16**, 840 ; 1949, *Ibid.*, **17**, 1198 ; 1951, *Helv. Chim. Acta*, **34**, 2371 ; 1954, *Z. Elektrochemie*, **58**, 219 ; 1954, *J. Chem. Phys.*, **22**, 2098 ; 1954, *Z. Naturforsch.*, **9a**, 989 ; 1956, *J. Chem. Phys.*, **25**, 293.
LENNARD-JONES, J. E., 1937, *Proc. Roy. Soc.* A, **158**, 280 ; 1949, *Ibid.*, **198**, 1, 14 ; 1952, *J. Chem. Phys.*, **20**, 1024.
LENNARD-JONES, J. E., and POPLE, J. A., 1951, *Proc. Roy. Soc.* A, **210**, 190.
LIEHR, A. D., 1957, *Ann. of Physics*, **1**, 221.
LONDON, F., 1937, *J. Phys. Radium*, **8**, 397.
LONGUET-HIGGINS, H. C., 1948, *Proc. Phys. Soc.*, **60**, 270 ; 1950, *J. Chem. Phys.*, **18**, 265.
LONGUET-HIGGINS, H. C., and BROWN, D. A., 1955, *J. Inorg. and Nucl. Chem.*, **1**, 60.
LÖWDIN, P.-O., 1948, *A Theoretical Investigation into some Properties of Ionic Crystals* (Uppsala : Almquist and Wiksell) ; 1955, *Phys. Rev.*, **97**, 1474, 1509 ; 1956, *Advanc. Phys*, **5**, 1.
LÖWDIN, P.-O., and SHULL, H., 1955, *Phys. Rev.*, **101**, 1730.
LÖWDIN, P.-O., and YOSHIZUMI, H., 1959, *Adv. in Chem. Phys.*, in the press.

McConnell, H. M., 1956, *J. Chem. Phys.*, **24**, 460.
MacDonald, J. K. L., 1933, *Phys. Rev.*, **43**, 830.
McGarvey, B. R., 1957, *J. Chem. Phys.*, **27**, 68.
McWeeny, R., 1951, *Proc. Phys. Soc.* A, **64**, 261 ; 1952, *Ibid.*, **65**, 839 ; 1954, *Proc. Roy. Soc.* A, **223**, 63, 306 ; 1957, *Proc. Phys. Soc.* A, **70**, 593.
Manning, P. P., 1954, *Proc. Roy. Soc.* A, **225**, 136, 244.
Margenau, H., and Murphy, G. M., 1956, *Mathematics of Physics and Chemistry*, 2nd edn (Princeton, N.J. : van Nostrand).
Melvin, M. A., 1956, *Rev. Mod. Phys.*, **28**, 18.
Moffitt, W., 1951, *Proc. Roy. Soc.* A, **210**, 245 ; 1954, *J. Chem. Phys.*, **22**, 1820.
Moffitt, W., and Ballhausen, C. J., 1956, *Annu. Rev. Phys. Chem.*, **7**, 107.
Mulliken, R. S., 1935, *J. Chem. Phys.*, **3**, 375 ; 1949, *J. chim. Phys.*, **46**, 497.
Murrell, J. N., and Longuet-Higgins, H. C., 1955, *Proc. Phys. Soc.* A, **68**, 329.
Pariser, R., 1956, *J. Chem. Phys.*, **24**, 250.
Pariser, R., and Parr, R. G., 1953, *J. Chem. Phys.*, **21**, 466, 767.
Parks, J. E., and Parr, R. G., 1958, *J. Chem. Phys.*, **28**, 335.
Parr, R. G., Craig, D. P., and Ross, I. G., 1950, *J. Chem. Phys.*, **18**, 1561.
Parr, R. G., Ellison, F. O., and Lykos, P. G., 1956, *J. Chem. Phys.*, **24**, 1106.
Parr, R. G., and Mulliken, R. S., 1951, *J. Chem. Phys.*, **19**, 1271.
Pauling, L., 1936, *J. Chem. Phys.*, **4**, 673 ; 1940, *Nature of the Chemical Bond*, 2nd edn (Cornell : University Press).
Pauling, L., and Wilson, E. R., 1935, *Introduction to Quantum Mechanics* (New York : McGraw-Hill).
Pitzer, K. S., 1953, *Quantum Chemistry* (New York : Prentice-Hall).
Platt, J. R., 1950, *J. Chem. Phys.*, **18**, 932 ; 1954, *Ibid.*, **22**, 1448.
Pople, J. A., 1950, *Proc. Roy. Soc.* A, **202**, 323 ; 1953, *Trans. Faraday Soc.*, **49**, 1375 ; 1958, *Mol. Physics*, **1**, 175.
Potts, R. B., 1953, *J. Chem. Phys.*, **21**, 758.
Preuss, H., 1956, *Integraltafeln zur Quantenchemie* (Berlin : Springer) ; 1957, *Z. Naturf.*, **12a**, 599.
Pullman, A., 1947, *Ann. Chim.*, **2**, 5.
Pullman, A., and Pullman, B., 1955, *Adv. in Cancer Research*, **3**, 117.
Ramsey, N. F., 1950, *Phys. Rev.*, **78**, 699.
Rittner, E. S., 1951, *J. Chem. Phys.*, **19**, 1030.
Roothaan, C. C. J., 1951, *Rev. Mod. Phys.*, **23**, 69.
Ruedenberg, K., 1953, *J. Chem. Phys.*, **22**, 1878.
Sahni, R. C., 1953, *Trans. Faraday Soc.*, **49**, 1246.
Schmid, L. A., 1953, *Phys. Rev.*, **92**, 1373.
Shull, H., 1959, *J. Chem. Phys.*, in the press.
Simpson, W. T., 1948, *J. Chem. Phys.*, **16**, 1124 ; 1949, *Ibid.*, **17**, 1218 ; 1958, *Ibid.*, **28**, 972.
Slater, J. C., 1929, *Phys. Rev.*, **34**, 1293 ; 1930, *Ibid.*, **35**, 210 ; 1951, *J. Chem. Phys.*, **19**, 220.
Sponer, H., and Teller, E., 1941, *Rev. Mod. Phys.*, **13**, 75.
Steensholt, G., 1947, *Phil. Mag.*, **38**, 748.
Titchmarsh, E. C., 1958, *Eigenfunction Expansions*, Part II (Oxford : University Press).
Watanabe, H., 1957, *J. Chem. Phys.*, **26**, 542.
Weinbaum, S., 1933, *J. Chem. Phys.*, **1**, 593.
Wheland, G. W., 1942, *J. Amer. Chem. Soc.*, **64**, 900.
Wilson, E. B., Decius, J. C., and Cross, P. C., 1955, *Molecular Vibrations* (New York : McGraw-Hill).

PROGRESS IN VACUUM TECHNOLOGY

By J. POLLARD

Services Electronics Research Laboratory, Baldock, Herts.

CONTENTS

	PAGE
§ 1. Introduction	34
§ 2. Mechanical pumps	35
2.1. Rotary backing pumps	35
2.2. Positive displacement rotary compressors	36
§ 3. Vapour pumps	38
3.1. Condensation pump theory	38
3.2. Oil condensation pumps	41
3.3. Mercury condensation pumps	42
3.4. Ejector pumps	43
§ 4. Vapour traps	44
4.1. Conduction cooled traps	44
4.2. Refrigerated baffle traps	45
4.3. Traps for oil-pumped systems	46
§ 5. Gettering and ionization pumps	46
5.1. Gettering pumps	46
5.2. Ionization pumps	51
§ 6. Measurement of pumping speed	52
§ 7. Measurement of pressure	53
7.1. Mechanical gauges	53
7.2. Acoustic gauges	55
7.3. Thermal conductivity gauges	55
7.4. Knudsen type gauges	56
7.5. Radioactive ionization gauges	57
7.6. Measurement of very low pressures	57
§ 8. Very low pressure techniques	62
§ 9. Newer mass spectrometers for vacuum systems	64
9.1. The omegatron	64
9.2. The linear R.F. spectrometer	66
9.3. The time-of-flight mass spectrometer	67
§ 10. Leak detection	67
10.1. Detectors using ordinary vacuum gauges	68
10.2. Detectors employing special properties of the search gas	68
10.3. Mass spectrometer methods	69
10.4. Performance of leak detecting systems	69
§ 11. Conclusion	71
References	71

Abstract. This article is concerned with advances in vacuum technology during the last two decades, although the achievements of the second half of this period have received most emphasis. The account is written from the point of view of the physicist and does not refer specifically to the considerable progress in the applications of vacuum technology to such specializations as metallurgy or chemical engineering.

In dealing with mechanical pumps some of the more recent improvements in rotary backing pumps are described, and the advent of the positive displacement rotary compressor is noted and its advantages discussed. The controversial subject of condensation pump theory receives brief attention, and sections on both oil and mercury condensation pumps give the more salient points of the advances leading to the present capabilities of these pumps.

Ejector type vapour pumps have been in use for a long period but the use of ejector principles at lower pressures is mentioned as a more recent development.

Trap design has kept reasonably in step with progress in vapour pumps, but unsolved physical problems bearing on the efficiency of traps are shown to remain.

New methods of pumping are challenging the established position of the vapour pump, and the achievements of various types of ionization and gettering type pumps are reviewed, as is some of the work on the mechanism of pumping in devices of this nature. This subject emerges as one holding many interesting problems, both theoretical and practical, and the progress in recent years is an encouraging pointer to the future.

Comparison of pumping speeds quoted by different workers has been difficult, due not only to differing definitions but to large errors inherent in many methods of speed measurement. Progress in understanding and resolving this problem is described.

In a section on pressure measurement progress in various types of gauges is assessed. Recently long-standing obstacles to measuring very low pressures have been overcome; the methods which have been developed, including some still in the laboratory stage, are discussed. The whole new subject of very low pressure techniques is one in which there has been rapid advance; this is surveyed and future trends are suggested.

Conventional deflection-type mass spectrometers have been omitted, but newer types which have found application as convenient tools in vacuum systems are described in principle. These include the omegatron, the linear r.f. types and the time-of-flight types.

The growth of leak detection as a branch of vacuum technology is discussed and a brief review of some methods is given. Reasons are advanced for the present difficulty of quantitative comparison of methods, and suggestions for progress are given.

§1. Introduction

Progress in vacuum technology has until recent years been less spectacular and perhaps slower than that in many branches of physics, but it is difficult by referring only to the last few years to give a review which is comprehensive and sufficiently broadly based for those not working in the field. In some parts of the subject it is necessary to refer to wartime research, and from the viewpoint of industrial vacuum technique the beginning of the present period of marked progress and expansion may perhaps be put at some fifteen years ago. Apparatus which before then was little used outside the research laboratory or the electronic valve industry is now employed in a wide range of industries. The size and range of pumping equipment available have greatly increased as a result of industrial demand, and auxiliary techniques such as leak detection have been vastly improved to deal with the increased size and complexity of the plant employed. The urgent demands of the atomic energy programmes contributed in large measure to this appearance in industry of equipment based on that of the laboratory, and, as might be expected, much of the progress has been practical in nature: scaling up in size and capacity of existing designs, improvements in detailed design and improved reliability are perhaps more evident than are advances in basic knowledge of the mode of operation of vacuum apparatus. Nevertheless, basic progress has also been made, the nature of some of the problems has been revealed and vacuum technology is moving out of the realm of mystic art into that of science.

In the laboratory new techniques are growing up in the field of very low pressures, and the progress here may well be already as important as that which has extended over a longer period in equipment for more normal pressures; already there have been useful results which would not have been obtainable at

higher pressures, and it is evident that all the avenues of possible progress have not been explored. Furthermore, completely new methods of pumping are appearing, and tending to make their way into industry with a speed which foreshadows many changes in established practice.

Vacuum technology is a field in which workers in various branches of science have been active, and it is not surprising that units and definitions have been confused and ambiguous. A step forward here has recently been made by the publication of a British Standards Glossary (B.S. 2951:1958) whose recommendations will be followed in this article.

§2. Mechanical Pumps

2.1. *Rotary Backing Pumps*

Rotary pumps of both the vane and the rotating plunger types have been used for many years, both for backing vapour pumps and by themselves where a comparatively rough vacuum is required. Details of the various designs are given in the textbooks, and modern types cover a range of displacements from 0·3 litre sec^{-1} to 300 litres sec^{-1}. It is generally known that a normal rotary pump is inefficient in pumping systems containing vapours, because the vapour condenses when the compression cycle reaches a point corresponding to saturation at the temperature of the pump. The resulting liquid is expelled through the discharge valve into the pump oil and can re-enter the oil feed and evaporate on the vacuum side of the pump.

A significant advance due to Gaede enables vapours to be pumped. Known as the gas-ballast pump (Gaede 1947), the principle is that a controlled amount of gas is admitted to the compression space through a one-way valve to join the vapour and gas already trapped in the space from the vacuum system. Compression proceeds until the pressure of the mixture is sufficient to open the discharge valve, but as the admitted gas provides most of the pressure needed to open the valve the partial pressure of the vapour is kept below saturation point, and the vapour passes out through the pump oil uncondensed. The rate of vapour intake possible depends on the amount of ballast gas admitted, and this cannot be increased indefinitely because the ultimate pressure attainable by the pump is worsened as the gas ballast is increased. A typical single stage pump of displacement 0·8 litre sec^{-1} is quoted as giving an ultimate pressure of the order of 0·006 mb unballasted, worsening to 0·6 mb with full ballast, at which it is capable of dealing with approximately 0·01 g sec^{-1} of water vapour. The effect of gas ballast of this order on the pressure–throughput characteristics of such a pump is shown in figure 1.

The maximum rate of vapour removal has normally been calculated as that amount which will produce a saturation value of partial vapour pressure at the known gas-ballast intake, but it has recently been reported that gas-ballast pumps have been operated without serious oil contamination at vapour intakes greatly in excess of the calculated figure (Power and Kenna 1955).

In their paper, which gives a useful discussion of the vapour pumping characteristics of gas-ballast pumps, these workers suggest that this may be due to the fact that a mist of condensed vapour is carried out of the pump with the ballast gas.

Where it is required to collect the pumped vapour a closed-circuit type of gas-ballast system has been used, the same ballast gas being re-circulated and the vapour removed by a condenser in the circuit. Condensers are also used between two rotary pumps, one being gas-ballasted, when very large amounts of vapour have to be handled.

The feasibility of pumping vapours has led to a demand for gas-ballast pumps to deal with corrosive vapours, and there have as a result been instances of corrosion of the working surfaces of the pumps. An attack on this problem and on the problem of gumming of the oil has been made by the use of oils containing inhibiting additives, following engine lubrication practice. Considerable success has been reported although the choice of additives is restricted to those of low vapour pressure, in order that they shall not be removed by the gas-ballast action of the pump.

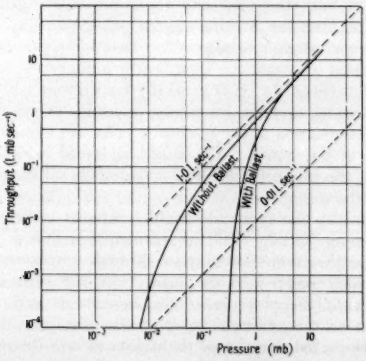


Figure 1. The effect of gas ballast on the speed characteristics of a rotary backing pump.

The noise and vibration of rotary pumps when operating at low pressures is familiar to all vacuum workers, and can be a considerable nuisance, particularly in the case of large pumps or when numbers are used in a confined space. Investigation of the cause has shown that it is mainly due to the sudden rapid acceleration of the discharge valve and the oil above it at the instant that the compression space becomes so small as to be completely filled with oil. Instantaneous pressures of the order of 100 lb in^{-2} may be reached. The result of this work is that almost silent pumps are now available, although exact details of the modifications made have not been published.

2.2. *Positive Displacement Rotary Compressors*

Until a few years ago the only way of obtaining high pumping speeds with mechanical pumps was to connect several rotary oil pumps in parallel. For mechanical reasons, such as unbalance of the rotor, there is a limit to the increase of size and of speed of rotation. Furthermore such pumps have low efficiency

due to the viscosity of the sealing oil, and this becomes more important as size increases. The need for a mechanical pump with balanced rotating parts and low friction has been met by the development of high-speed, positive-displacement rotary compressors. The type most used is based on the principle of the Roots blower, shown in figure 2. Two long figure-eight shaped rotors are driven in

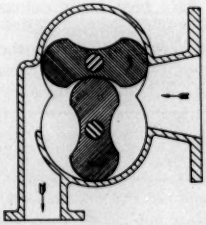


Figure 2. The principle of the positive displacement rotary compressor.

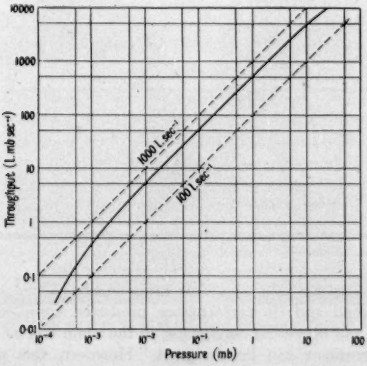


Figure 3. Speed characteristics of a typical positive displacement rotary compressor.

counter-rotation, touching neither themselves nor the pump wall, so that they are able to run unlubricated. Accurate machining is necessary to maintain close gaps to minimize back-leakage, clearances of 0·006 in. to 0·010 in. being typical. Rotor speeds in the range 1500 to 3000 rev min^{-1} are used. Although such pumps are able to exhaust directly to atmosphere, they are not frequently used in this way, but are backed by a rotary oil pump. At high pressures the power required to drive the rotors, converted to heat almost entirely inside the blower, can cause expansion and seizure unless precautions are taken, such as cooling the rotors or

in other ways which have been discussed in the literature (Nöller 1956, Thees 1955). Moreover, the compression ratio falls off at high pressures, and is as low as 3 for exhaust to atmosphere compared with values of 40 to 50 for the range of pressures between 10^{-3} mb and 1 mb, so that this type of pump has found its main use in the latter region, particularly as its speed is well maintained down to pressures below 10^{-3} mb (figure 3), enabling a vapour pump to be dispensed with for many purposes. There have been many examples of its usefulness in this respect in vacuum metallurgy, where pumps of this kind with speeds up to 800 litres sec^{-1} have been used extensively. A further advantage claimed for use in the metallurgical field is that dry dust passes harmlessly through the pump.

Recent publications give considerable detail on the characteristics of positive displacement blower pumps used in various ways, and discuss the effect of such parameters as backing pump speeds on overall performance (Van Atta 1956, Steinherz 1956).

In some vacuum applications such as mass spectrometry the vapour of the backing pump oil is a nuisance in the system, and small positive displacement blower pumps may have a future application here, as they can be backed by water jet pumps. Present designs cannot be said to be oil-free, because of the presence of lubricated bearings or shaft seals, but this is a problem which should not be difficult to solve.

§3. Vapour Pumps

3.1. *Condensation Pump Theory*

The theory of these pumps has, from their earliest days, been a controversial subject, and a large literature of varied contributions, not all helpful, has accumulated. Any account of progress in this field is likely to be in itself controversial, and to give any semblance of completeness would need more space than the whole of this article ; but in discussing progress in the techniques of vacuum, briefness on this subject is justified, since practical progress has occurred in spite of the state of the theory.

It is scarcely an exaggeration to say that after forty years, condensation pumps are designed on a basis of practical experience and guesswork, enlightened guesswork though it may be. There is still no quantitative theory which allows a pump to be designed and its performance predicted completely from a knowledge of its design and operating conditions, in the same way as, say, a transformer or an optical instrument can be designed. However, this position represents progress, since for many years the Gaede diffusion theory was accepted as applying to all pumps, and exercised a stranglehold on condensation pump design. Modern adherents to this theory have to struggle to modify and extend it to conform to the performance of existing pumps (Jaeckel 1947, Vogt 1954). The inadequacy of the model taken by Jaeckel has already been pointed out by Lifshitz and Rosensweig (1952) who themselves put forward, as a beginning, an examination of the simple case of a plane parallel uniform stream of vapour condensing completely on meeting a wall at an angle. Their theoretically derived isobars are of the same form as those obtained experimentally by Alexander (1946) and the reason for the main difference is discussed by them.

The work of Alexander marks a definite step forward in condensation pump theory, as well as in the practical design of pumps. His 1946 paper pointed out the pitfalls in applying Gaede's theory to pumps other than the original Gaede diffusion pump, and suggested different design principles, which led to successful high-speed mercury vapour pumps. Among other principles, Alexander showed the desirability of using a high-speed jet of vapour of adequate density, and used

Figure 4. Detached shock wave in front of spherical obstacle in flow of mercury vapour from a pump nozzle.

convergent–divergent turbine type nozzles to attain this object. He considered the proportion of molecules backstreaming from the jet in a direction opposed to the pumping direction as a function of the ratio of vapour streaming velocity to most probable random velocity, and gave experimental values of this ratio between 1·2 and 3·4. His values of most probable random velocity were calculated assuming a vapour temperature equal to that at the entrance to the nozzle, and are therefore incorrect, as they take no account of the considerable drop in temperature due to the adiabatic expansion through the nozzle. This inaccuracy did not affect the correct conclusion—that the streaming velocity of the vapour can be considerably higher than the most probable random thermal velocity.

At that time it had not been realized that, for boiler pressures above a critical value, the vapour flow from such pump nozzles must reach supersonic velocity. This was demonstrated experimentally by the present author at the Physical Society's Exhibition of 1950, by showing shock waves formed by obstacles in the flow of vapour from a mercury pump nozzle, the vapour being excited by a radio-frequency discharge (see Barrell 1950). Figure 4 shows a typical photograph taken on this apparatus, which was described in the following year (Pollard 1951). Experimental confirmation has come from, *inter alia*, Klumb *et al.* (1953) and Nöller and Kutscher (1955).

Nevertheless, modern textbooks still contain statements such as "... streaming velocities that are large in comparison with thermal velocities cannot, in fact, be generated by a process in which a vapour is accelerated by its own expansion, whether in a cylindrical or in a divergent nozzle" (Reimann 1952). Let us look at this statement in the light of the fact that for flows from mercury pump nozzles Mach numbers of 2·0 are easily demonstrated, i.e. $v/c = 2$ where v is the streaming velocity and c is the local velocity of sound. If the most probable thermal velocity is u, then it may be shown that $u = c(2/\gamma)^{\frac{1}{2}}$ where γ is the ratio of specific heats for the vapour, in this case 1·67. Then $(v/u)\,(2/\gamma)^{\frac{1}{2}} = 2$, so that $v/u = 1{\cdot}93$. Here, then, the streaming velocity is nearly twice the most probable thermal velocity. This may not seem physically so improbable if one calculates the effective vapour temperature after expansion, which in the case given is -71°c, taking a boiler temperature of 200°c. For a Mach number of 3 it would be -155°c.

It can be taken as established that, in general, the flow from vapour pump nozzles is supersonic, and that any theory which does not take the effects of this into consideration is likely to meet considerable difficulties, to put it mildly. One result which may be mentioned concerns the relation between the γ for the vapour and the spread of vapour from a nozzle: it becomes evident that a nozzle can be designed for a high γ vapour such as mercury which will give negligible backstreaming from the vapour jet, but that this is impossible for a low γ fluid such as an oil vapour.

Perhaps the fullest recent attempt at a complete theory, taking the fact of supersonic vapour flow into account, is that of Nöller and Kutscher (1955). The theory predicts the form of pumping speed curve for condensation pumps and for ejectors, and gives a satisfactory treatment of critical backing pressure. The main points of this work have been summarized by Nöller (1955). The theory links pumping speed to density of jet vapour, higher density giving less penetration by gas and lower speeds. This is open to criticism, as work at this laboratory has shown in many pumps that the speed can be independent of boiler pressure over a wide range as long as the cooling of the condensing wall remains adequate. The conditions on the condensing wall, both the temperature and the nature of the condensation, have been shown to have a large effect on speed and this aspect is ignored completely by Nöller. So also is the effect on speed of the impedance of the vapour filling the pump above the pumping region, which, as has been shown by Stark (1957), can be appreciable in the case of mercury. Nöller's implication that modern improvements in pumps have been obtained by deliberate use of supersonic vapour speeds should not go unchallenged. It would be much nearer the truth to say that the realization that Gaede's theory applied only to his

original design of diffusion pump has led to removal of restrictions in the mouth of the pump with corresponding improvements in speed.

3.2. *Oil Condensation Pumps*

The oil condensation pump has shown little obvious change during two decades except in size. Scaling up has proceeded to the point where speeds of tens of thousands of litres per second are not uncommon. The modern pump is an advance on its predecessors in other ways, however, independent of size.

Better understanding of the fractionation process has led to the use of self-purifying pumps and the incorporation of self-purification in fractionating pumps. It had been realized before 1945 that fractionating pumps did not always give the results expected. To obtain pressures as low as the vapour pressure of the pump fluid it was necessary to use a horizontal type 3 or 4 boiler pump (Hickman 1940) under laboratory conditions. Under less rigid conditions these pressures were often not obtained, and the concentric-boiler type fractionating pump (Malter and Marcuvitz 1938) often gave pressures worse than ten times the vapour pressure figure. Hickman (1951) showed that the relative volatilities of mixtures of liquids increase at low rates of distillation, making it harder than had been supposed to fractionate decomposition products and higher vapour pressure impurities in pump fluids. Experiments showed the composition of condensate to be a function of position in the pump, the lightest constituents moving preferentially to the low pressure side. For instance, Octoil (2-ethyl-hexyl-phthalate) contaminated with less than 1% of butyl phthalate and used in a three-stage fractionating pump was found to pass vapour through the low pressure entrance as if it had contained 27% butyl phthalate. The modern self-purifying pump is a partial answer to this situation because it causes the volatile constituents to be removed from the pump instead of being returned to any boiler compartment. Various designs have appeared, the most common one being arranged so that condensed liquid returning to the boiler is passed over a hot surface in the backing space, the volatile products of this 'falling film' distillation passing out of the backing side. This method has been used alone (Latham *et al.* 1951) and in conjunction with concentric-boiler fractionation. Hickman and Kinsella (1956) have described a single-boiler pump using this principle in conjunction with a reservoir for intermediate pressure volatiles which do not easily pass out of the pump. The pump is provided also with a reservoir into which the liquid charge is forced by atmospheric pressure whenever the pump is opened to atmosphere, the design being such as to give a small area of contact between liquid and air. This, together with heaters of low heat capacity, enables very rapid warming up and shutting down cycles to be achieved, even with conventional oils such as 'Octoil', without the need for such expedients as boiler cooling coils. It should be mentioned that similar rapid cycles have also been employed satisfactorily in this country with normal commercial oil pumps using the silicone fluids.

Much progress has been made in overcoming another drawback of oil condensation pumps—backstreaming, which is defined as the flow of pump vapour, however caused, through the mouth of a pump in a direction contrary to the pumping direction. The ratio of specific heats for oil vapours being low it is impossible to design a practical simple oil pump nozzle from which vapour will

not spread in a backward direction. The obvious way of attempting to deal with this is to place a cooled skirt round the nozzle so as to intercept and condense vapour emerging at angles greater than the maximum required, and this has been successfully incorporated in commercial pump designs (Power and Crawley 1954). Other less fundamental causes of backstreaming were also investigated by these workers, who describe alterations in pump design features to eliminate or reduce them.

Backstreaming in oil pumps has not been reduced to the point where no baffle is required for many applications and such points as exposure of hot surfaces, e.g. vapour pipes, to oil inside pumps have not been seriously attacked. The work which has been done has effected a great improvement, however, and pumps can be obtained with backstreaming rates as much as fifty times lower than their counterparts of even a few years ago.

Progress in the oils themselves has not been spectacular, and in the case of the hydrocarbons and the esters is confined largely to more careful purification. Our knowledge of the lowest pressures obtained with different oils in pumps has improved, following the work of Blears (1947). This showed that when the errors due to vapour absorption by the ionization gauge were eliminated the lowest pressures obtainable with an oil pump alone proved to be much higher than had previously been believed.

Since their original introduction (Brown 1945) the silicone oils have been extensively used and have proved their value when high resistance to decomposition on exposure to air is required (Latham *et al.* 1952). They have not been trouble-free, however, and have tended to produce very tenacious insulating films under some circumstances of exposure to hot surfaces or electrical discharges. It has been claimed that some of these effects have been due to impurities in the earlier samples which have now been eliminated. Another criticism, of variability of characteristics between batches, may be due to the fact that most of the silicone oils used have been residues from other processes. The latest one, which is only just becoming commercially available, is manufactured as a pump oil, and appears to have advantages over the previously used silicones. Its vapour pressure is lower and it is less improved by fractionation, indicating a more homogeneous composition with a lower proportion of higher vapour pressure constituents.

The use of glycerol as a pump fluid has also been reported (Alexander 1948, Pollard *et al.* 1948). It has advantages for some purposes, being cheap, non-toxic and easily removed from surfaces, and not showing decomposition into tarry substances or conducting or insulating films; its decomposition products under normal use in a pump appear to be gaseous. Its vapour pressure involves the use of a cold trap, but it is more easily trapped than mercury. It is perhaps better to look on it as an alternative to mercury rather than to the oils, and apart from nozzle design, mercury pump techniques can be applied to it.

3.3. *Mercury Condensation Pumps*

In spite of improvements in oil pumps the mercury pump has remained in use, finding applications where even small amounts of oil vapour are detrimental. These include work with mass spectrometers as well as on thermionic and photo-emission devices and in electronic apparatus operating at high voltages, where

the deposit of oil or cracked products on insulators can cause breakdown. Mercury pumps have become popular also for pumping positive-ion accelerators, where in addition to its normal disadvantages in high voltage systems oil has the further drawback of a higher collision cross section per molecule than mercury. Not only does this increase the proportion of beam lost at a given pressure, but produces a greater number of electrons. These are accelerated and produce x-rays which ionize the high pressure gas insulation causing erratic voltage distribution or breakdown.

This accelerator application has been the main reason for the recent development of large mercury condensation pumps. For most of the other applications speeds up to a few hundred litres sec^{-1} are adequate, but accelerator requirements have resulted in mercury pumps being available up to 32 in. diameter with untrapped speeds reaching 14 000 litres sec^{-1} (Smith 1955). These are based on oil pump designs, and could probably be more efficient, though some note has evidently been taken of the work of Alexander (1946). In smaller sizes, at least, the belief that mercury pumps must be slower than oil pumps of corresponding size has been shown to be wrong. Mercury pumps have to be trapped, of course, and it is the trapped speed which is of interest. Advances in traps will be discussed in § 4 ; there seems to be no fundamental reason why mercury traps for normal pressures should have much greater impedances than the corresponding oil pump baffles.

3.4. *Ejector Pumps*

The ejector pump, using steam or compressed air, was in use long before Gaede's original diffusion pump experiments, and it is unfortunate that the ejector engineer and the vacuum physicist have remained so far apart for nearly half a century. To those familiar only with condensation pumps it came as a surprise to find multistage steam ejectors maintaining process plant at pressures as low as 10^{-1} mb without traps, in spite of a vapour pressure of water of 20 mb at condensation temperature, and at the same time exhausting to atmospheric pressure. Ejector principles have begun to find their way into vacuum technology in recent years, and so-called oil ejector and mercury ejector pumps are available, finding their main application where large throughputs have to be handled at pressures from a few millibars down to below 10^{-2} mb—a region between the effective ranges of rotary oil pumps and of condensation pumps (but which is being invaded by the positive displacement blowers (§ 2.2)). Oil ejectors with speeds up to 5000 litres sec^{-1} and critical backing pressures of the order of 10 mb are available, and mercury ejectors pump against backing pressures above 100 mb. The demarcation between such pumps and the older type of condensation pump is not at all clear cut. Some have many ejector features but condense their operating fluid on the diffuser wall. Some are little more than conventional condensation pumps working with a fluid suitable for high boiler pressure operation and designed to work in a higher pressure range than usual and with a higher critical backing pressure ; these advantages are usually obtained at the expense of speed at the lower pressures. These pumps are often termed 'boosters' as they can be used to back condensation pumps when the throughput or backing requirements preclude the use of the latter alone but the low pressure requirements are beyond the range of the 'booster'. The term is a bad one, however,

as it is used in the opposite sense in steam ejector practice. Here a booster is the first pump on the high vacuum side of a multistage ejector system, where it 'boosts' or compresses steam from a low pressure to a higher pressure, at which it is possible to condense it at the temperature of the cooling water available. Confusion will not be lessened by the tendency on large plants to back oil ejectors by stages of steam ejector pumping where steam is available and running cost considerations favour composite systems.

Confusion of terminology does not, however, obscure the fact that practically the whole of the normally required pressure range is now covered by vapour pumps, and combinations can be assembled to suit a wide range of throughputs in a variety of applications.

§4. Vapour Traps

4.1. *Conduction Cooled Traps*

With the development of high speed mercury vapour pumps the design of cold traps had of necessity to progress from the types which had been immersed in a Dewar flask. The early high-conductance types reversed the procedure and

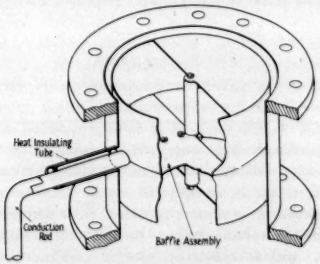


Figure 5. Conduction-cooled trap.

placed the coolant inside the vacuum system by using concentric spheres or cylinders, the inner being filled with coolant through a heat-insulating tube passing through the outer wall. These traps, while giving an improvement in conductance, had numerous disadvantages which are overcome in the conduction-cooled baffle type (Pollard 1953). In the original design of this type a baffle assembly consisting of three semi-circular copper sheets spaced along a copper cylinder was cooled by conduction along a copper rod, thermally insulated from the wall by a thin-walled stainless steel cylinder and dipping into a Dewar flask outside the system. A somewhat later design of larger size is shown in figure 5. It is $5\frac{1}{4}$ in. in diameter and has a conductance of about 175 litres sec^{-1} for air. The hottest part of the baffle cools to $-130°c$ in 20 minutes and $-150°c$ in 30 minutes after applying liquid oxygen to the rod (ambient temperature 20°c). For a given size of trap there is an optimum proportion of dimensions for the parts of the baffle assembly. Too high a thermal capacity gives a slow cooling rate and reduction

in thickness in the wrong places gives a high final temperature due to inadequate conduction. Martin (1955) gives an account of an unsuccessful attempt to design a trap of this type only slightly larger than that of figure 5, which indicates difficulties of this sort. His final more complicated design where the coolant is led near to the baffles from an external reservoir would probably be worth while for much larger sizes, however. The baffle design is, of course, optically opaque, so that no molecule can pass through without at least one collision with a cold surface. There are other geometries equally suitable for this purpose, and for a given outside size of trap there is not much to choose between them from the point of view of conductance. Typical ones which have been used are sketched in figure 6.

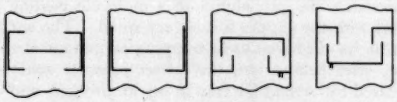


Figure 6. Optically opaque cold trap geometries.

This 'optically opaque' type of design has also been applied more recently in conjunction with an internal coolant reservoir. In the larger sizes, if sufficient care is taken over heat insulation of the reservoir these internal reservoir traps have the advantage of long intervals between filling. Disadvantages are that they cannot quickly be warmed up as can conduction types and are more difficult to de-gas and to dismantle for cleaning.

4.2. *Refrigerated Baffle Traps*

When even larger pumping systems, in the 1000–10 000 litres sec^{-1} range are considered traps directly cooled by such coolants as liquid oxygen become impracticable, and designs of refrigerated baffle type traps have recently been reported from various quarters. Two-stage mechanical refrigeration is possible down to $-100°C$ and particulars of such a system operating on a 32-in. pumping unit have been given by Kennedy and Smith (1955). There is a lack of information on the temperatures required for trapping mercury. Textbooks still assume that the pressure of mercury vapour attainable above a trap is equal to the vapour pressure of mercury at the trap temperature, but this has been shown not to be the case by several workers. Although Knudsen's conception of a critical temperature below which mercury was condensed efficiently and above which condensation did not occur is no longer accepted, there is no doubt that low values of condensation coefficient are common for mercury; data for its variation with temperature on surfaces of different materials are not available, however.

For conduction traps with baffles of clean copper, nickel-plated copper and chromium-plated copper it has been found at this laboratory that when solid CO_2 is employed as a coolant the pressure of mercury vapour above the trap cannot be reduced much below 10^{-5} mb, although the corresponding vapour pressure is 500 times smaller than this. Results of this nature point to the need for more information before refrigerated traps for mercury can be designed with confidence.

4.3. *Traps for Oil-pumped Systems*

All types of cold trap and refrigerated baffle have been used with oil pumps where the possibility of oil vapour contamination has had to be avoided. Oil vapour has also been trapped with baffles of activated charcoal, following the work of Anderson (1937) and considerable success has been reported with traps of this type. The most interesting development in non-cooled methods of trapping oil vapours is the copper trap used by Alpert (1953). A corrugated copper foil (0·003 in. thick) is rolled to form a cylinder filling part of the tubing of the pumping line. The corrugations are of such a size that the folded foil produces the effect of filling the tube with a large number of smaller tubes about 0·05 in. diameter. The surface area of copper is therefore relatively large and at the relevant pressures the probability of a molecule passing through without numerous collisions with the copper foil is very small. The surprising fact about the trap, however, is its effectiveness in trapping oil vapour at room temperature. Alpert found that, after baking out, the lowest pressure achieved above such a trap was not decreased by cooling the trap in liquid nitrogen, and the low pressures attained were maintained for periods of a week or more. There has been no satisfactory explanation of oil-retaining efficiency of this order, and experiments to investigate saturation conditions have not yet been reported. With its present geometry the trap is limited to fairly low-speed systems, but the possibilities if it could be shown to be efficient with fewer collisions per molecule with the copper surface are obvious.

§5. Gettering and Ionization Pumps

5.1. *Gettering Pumps*

Apart from the use of getters in sealed-off devices numerous gettering procedures have been used for removal of gas in vacuum apparatus. These include such methods as evaporation of barium and similarly reactive elements, as well as the use of solid getters such as zirconium run at appropriate temperatures. Many of these are described in the textbooks, Dushman (1949) giving numerous examples and references. The design of pumping apparatus using continuous evaporative gettering is a more recent step forward, the first work of note being that of Herb *et al.* (1953) leading to the Evapor–ion pump whose design and characteristics were described by Davis and Divatia (1954). The gettering action is produced by the evaporation of titanium on to the walls of the pump, this metal being chosen because of its chemical affinity with the common gases and because it and its compounds are stable and have low vapour pressures at room temperature. Figure 7 gives a diagrammatic representation of the essential points of one of the early designs. Titanium wire is fed from reel R down a guide and evaporates on contact with a post P of graphite or of metal alloy which is heated to about 2000°c by electron bombardment from filament F, maintained at 1 kv negative. For a pump 12 in. in diameter speeds of about 1800 litres sec^{-1} for nitrogen and 3000 litres sec^{-1} for hydrogen have been reported, at evaporation rates of about 5 mg of titanium per minute. The considerably higher values of speed quoted by Davis and Divatia are of doubtful validity due to their method of speed measurement, their leak being led into the middle of the pump and the

pressure measured at a remote point (see §6). The titanium surface does not getter inert gases and ionization of these is effected by including an accelerator grid A taking emission from the filament and repelling to the wall all ions formed outside it. When ionized the inert gases are pumped, speeds in the original pumps being only a few litres sec^{-1} for argon for accelerator currents of the order of 200 milliamps.

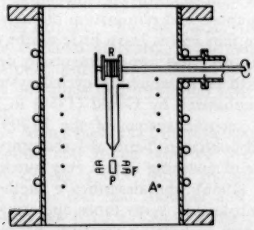


Figure 7. Representation of the essentials of an Evapor–ion type gettering pump.

Before starting the pump a pressure below that at which discharges can occur must be provided, and in early pumps this was achieved by a trapped backing pump in conjunction with a six-hour bake to 300°c. Later pumps were provided with means of rapidly evaporating an auxiliary charge of titanium, thus avoiding the need for the bake. Details of the performance of a commercial pump of this type have been given by Swartz (1955) who also discusses the important question of life of the various components. Swartz's results throw light upon the gettering efficiency by showing that for nitrogen at a typical evaporation rate the maximum speed was attained at a combination ratio of 4 (i.e. one nitrogen atom pumped for every four titanium atoms evaporated). The pumping speeds for nitrogen and hydrogen decreased markedly with increase in wall temperature, but those for argon and water vapour were independent of temperature in the range 25°c to 100°c.

Alexeff and Peterson (1955) describe an experimental Evapor–ion pump designed for the study of the problem of pumping the inert gases. They have achieved a significant advance with a simple grid filament ionization system with a high transparency grid and in the assembly of which all parts capable of absorbing electrons are either shielded or made very small. Such a system with an ionizing current of 1 amp gave a speed for argon of 120 litres sec^{-1} in the region of 10^{-5} mb. With an axial magnetic field speeds of 250 litres sec^{-1} were reached for argon and 90 litres sec^{-1} for helium. Measurements of pumping speed and ion current to the wall enabled the sticking probability to be measured and its variation with ion energy investigated. At an energy of 800 ev 25% of ions striking the wall remain permanently ; this figure drops to 20% at 500 ev, and 12% at 350 ev.

The pumping speeds for the inert gases obtained by these workers indicate that one of the disadvantages of the earlier pumps has already been conquered. Other disadvantages are rapidly being dealt with : titanium attack and build-up

on the post will be a trouble of the past when evaporation by direct electron bombardment of the wire as it emerges from a cooled guide is adopted; rapid pump-down after admission of air will be facilitated by initial reversal of polarities so as to heat the wall by electron bombardment; and the firing of starting primer charges will be eliminated by arranging for higher rates of evaporation to be available, which may be reduced as throughput requirements allow. Values of 0·5 gramme per minute have been used experimentally.

Its advantages of high speed and elimination of vapour contamination without cold traps, baffles or large flap valves leave little doubt that the Evapor–ion type pump has an important future, and improvements in design may well extend its use outside its present field of infrequently cycled systems such as accelerators. Its importance here is emphasized by Gould (1956) in his interesting discussion of the proposals for the vacuum system of the 30 000 MeV proton synchrotron being constructed at the Brookhaven National Laboratory.

An interesting example of pumping with a very much smaller getter–ion pump has been given by Gale (1956) who describes a method of maintaining a low pressure in cold-sealed, unbakeable, x-ray tubes operating at potentials up to 2 MV. The tubes are pumped down to backing pressure, processed, and then sealed off with their getter–ion pump which consists simply of a small chamber containing the evaporators (tungsten filaments overwound with titanium wire) to which is connected a commercial Penning gauge. By evaporating some titanium with the Penning gauge operating the pressure is reduced to below 10^{-6} mb. Pumping speed is low, figures quoted being 18 ml. sec^{-1} for nitrogen and 9 ml. sec^{-1} for helium. A point of interest is that the ionization region is separate from the evaporating chamber and cannot 'see' the titanium surface. The tentative explanation of inert gas pumping in the Evapor–ion pump, that ionized atoms are accelerated and driven into the wall, is thus not applicable in this case; it has been confirmed that inert gas pumping still occurs when an atom must make at least four collisions with walls before reaching the titanium surface. This may indicate that it is the excited state of the atom which is important, as it is known that these states are relatively long lived.

The work of Alpert, in which he used an ionization gauge to pump down to very low pressures (see § 8), together with the advent of gettering pumps, has stimulated renewed interest in the pumping action of ionization gauges. The removal of certain gases by a hot tungsten filament has been known and used for many years, Langmuir having published several papers on the subject in 1913. Oxygen is removed by a chemical reaction forming tungsten oxide, while hydrogen is dissociated into the atomic form which is actively adsorbed on the surfaces of the gauge. Pumping speeds for these processes agree with expected values, depending on the area of hot filament, values up to 1 litre sec^{-1} being measurable in conventional gauges. The other pumping effect, which requires the electrode voltages to be applied and causes the pumping of the inert gases is not as well understood, although it has in recent years been effectively employed as is illustrated by the work of Alexeff and Peterson previously referred to. Bloomer and Haine (1953) described experiments showing that the pumped gas was held by the glass walls. When the walls were shielded or coated with a colloidal graphite film no pumping occurred. For helium, results contradictory to this were

obtained by Varnerin and Carmichael (1955 a); their results linked the pumping of helium to the presence of a metallic film on the glass wall. After successive high-temperature degassing of the electrodes, resulting in increases of metallic film thickness, very significant increases in helium pumping speed were observed, but this dropped by a large factor after acid cleaning of the walls. These workers also showed that much of the helium pumped went into the filament supports.

Their results were confirmed by Young (1956), who found that a metal film improved the speed for helium but not for nitrogen. The speed for nitrogen was unchanged at about 0·1 litre sec^{-1} for his gauge when the wall was covered with a film of metal or of colloidal graphite. This value is considerably higher than that estimated by Alpert, but Young showed that the ion current to the walls could be as much as ten times that to the ion collector, and Alpert's figures were based upon the latter.

Most investigators have found inert gas pumping speeds to be proportional to ionizing current, but Bloomer and Haine reported this to be the case only for very small currents except during a short initial pumping period. Some of the discrepancies may be due to the fact that Bloomer and Haine did not work with definite gases, but pumped down from air, so that the gas composition may have been progressively changing, and at the lower pressures would in any case have been unknown. Even so, not all the differing results available are reconcilable, even qualitatively, and more work is obviously needed. In the case of helium a theory of Varnerin and Carmichael (1955 b) explains the reduction of pumping speed as more gas is pumped in terms of re-emission of helium previously pumped. Quantitative agreement was obtained with the results of pumping experiments in which pressure reduction from 10^{-4} mb to 4×10^{-8} mb was accompanied by a reduction in pumping speed by a factor of 400. This theory has been supported by further work by these authors (Varnerin and Carmichael 1957) on the adsorption of helium ions on a molybdenum surface. Hobson and Redhead (1958) agree with this explanation for helium and also quote measurements with nitrogen showing two similar pumping speeds, one for ionic pumping (speed proportional to ion current) and another for 'chemical' pumping, which do not appear to be independent of one another.

The Penning cold-cathode gauge has long been known to act as a gettering pump, and with normal electrode materials pumping speeds up to 0·5 litre sec^{-1} have been reported for air. The Redhead magnetron gauge (§ 7.6) has a speed of 0·13 litre sec^{-1} for helium, maintained down to very low pressures.

The combination of the Penning gauge principle with sputtered deposition of gettering material is the basis of the latest development in ionization–gettering pumps. This very significant advance has been described by Hall (1958). The main constructional features are shown in figure 8. Plane cathodes of 0·125 in. titanium sheet are contained in a stainless-steel box about 4 in. square, and enclose a subdivided anode of special design, consisting in effect of a mesh of open-ended boxes. A permanent magnet provides a field of 1200 oersteds, and a power supply gives up to 3 kv for the anode potential.

In operation the system is pumped down to a pressure of the order of $2{\cdot}5 \times 10^{-3}$ mb with a rotary pump, and the power supply switched on with the rotary pump still pumping. The power supply has filament lamps as limiting

resistors, a typical value of limited current at this stage being 140 mA. The discharge causes sputtering of titanium from the cathode, so that the large surface area of the anode is continually covered with a fresh layer of titanium. At this point the anode potential is about 400 volts: as the discharge and sputtering continue the pressure drops, due to the gettering action, current falls and the anode potential rises to the full 3 kv. The tap leading to the rotary pump is then closed and the 'electronic' pump continues to pump. The speed is of the order of 10 litres sec^{-1} for air, and it is noteworthy that the speeds for hydrogen, argon and helium differ little from this figure. Power consumption is about 3 watts at 10^{-5} mb, and below 10^{-6} mb amounts to milliwatts only, so that, after starting,

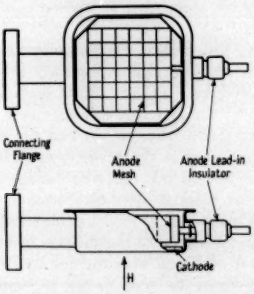


Figure 8. Construction of Penning-gauge type ionization–gettering pump. (*After* Hall.)

a number of pumps can operate from one power supply. Current–pressure calibration is consistent and stable so that the pump provides a pressure gauge also. The pump continues to operate at pressures below 10^{-9} mb, but speeds have not yet been measured below about 10^{-7} mb.

Sputtering by the discharge is an economical way of using the titanium, only as much being sputtered as is necessary to continue to pump the gas present. For such gases as air, hydrogen and carbon dioxide between one and two titanium atoms are sputtered for each gas atom pumped. The life of the pump should therefore be long, particularly when used for pumping systems with low gas evolution. A figure of 10 000 hours at a pressure of 10^{-6} mb is given as an estimate. Hall states that saturation effects need more investigation, but that it has been found that a pump whose ultimate pressure and ease of starting have deteriorated after long service can be re-conditioned by a bake under vacuum.

Applications in which this pump can conveniently replace vapour pumps and traps are obvious, and its impact will very soon be felt throughout the low and medium speed range. For higher speeds the disadvantage of providing a large magnetic field may well outweigh the advantage of sputtering as a means of deposition, and the evaporation type of gettering pump may be preferred.

Hall reports experiments with other cathode materials which showed the superiority of titanium from the point of view of pumping speed. Copper was

used with the aim of differentiating between true gettering, or chemical bonding, and the physical covering of pumped gas layers with fresh sputtered metal layers, and it is very interesting that the speed with copper was many orders of magnitude less than with the reactive metals, which included magnesium and molybdenum. This result holds even more interest when it is remembered that the inert gases argon and helium are pumped at approximately the same rate as air. It is very unlikely that they can be pumped by the same reactive mechanism, but if they are removed by some such means as the driving of ions into the cathode the problem of why this same mechanism does not apply to all gases when the cathodes are of copper immediately arises.

This pump is yet another example of a vacuum device which has been taken to a state of adequate practical efficiency without a satisfactorily complete explanation of its mode of operation in accord with experimental results. Similar problems remain to be solved in the whole range of gettering pumps, and the subject is one which offers both theoretical interest and the possibility of improved pumping devices following a better understanding of the gettering processes.

5.2. *Ionization Pumps*

Apart from the use of ionization in gettering type pumps, to increase efficiency or to bring about the removal of gases which otherwise would not be gettered, there have been many attempts to produce a true pumping action by ionizing gas

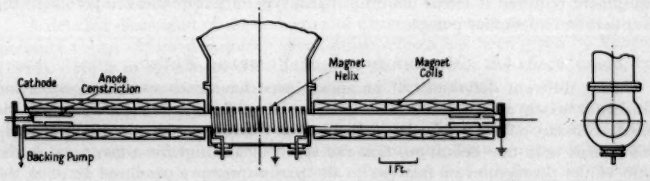


Figure 9. Simplified diagram of ionization pump. (*After* Foster.)

molecules and moving the ions to a higher pressure region by suitable field configurations. Most of the attempts have met with little success and those which have worked have been very slow. One design has been described, however, which has been taken to a quite different order of speed and efficiency (Foster 1953, Foster *et al.* 1953). It is shown diagrammatically in figure 9. The ionization is produced by a Penning type discharge, but supplied with electrons from a hot cathode at one end of a long tube which forms the anode. The cathode at the other end is cold, and the discharge is confined by a magnetic field of up to 2000 oersteds, produced by external windings on the tube and continued through the central section, which is the gas inlet, by use of a water-cooled open helix. The backing pump is connected to the hot cathode end, single-ended backing having been found to be sufficient.

In operation, with cathodes at -300 volts, a discharge extends from cathode to cathode, discharge current being 10 to 20 amps. Positive ions formed in the

discharge have a low velocity compared with the electrons, and a short mean free path. A positive space charge thus exists in the column, and the space charge field moves the positive ions towards the cathode. In this way the pressure in the central region is reduced until an equilibrium is reached between the number of molecules being ionized and removed and the number returning, by such processes as diffusion of ions from the discharge across the magnetic field to the anode or recombination in the discharge. On reaching a cathode some ions are gettered by the cathode material, others having been neutralized at the cathode are removed into the backing system. Other neutral molecules on leaving the cathode are gettered on wall surfaces facing the cathode, which are continually being renewed by metal sputtered from the cathode. The length of the pump is made sufficient that the probability that a neutral molecule can travel back to the centre without being ionized is very small.

As pressure is reduced the positive ion supply becomes too small to maintain an adequate discharge, and it was found necessary to introduce a leak of gas at the hot cathode end to keep the pressure there above 4×10^{-4} mb.

The lowest pressure attained at the pump mouth was of the order of 10^{-6} mb, and pumping speeds varied from 3000 litres sec^{-1} to 7000 litres sec^{-1} according to the nature of the discharge and the gas being pumped. The pump used a very high power input to achieve these results, however, the magnet winding alone taking 32 kw and the cathode 4·5 kw apart from the 6 kw of the discharge. For this reason, together with the large size and considerable amount of auxiliary equipment required it seems doubtful if this type of pump will compete with the evaporation and similar pumps.

§6. Measurement of Pumping Speed

Many different definitions of pumping speed have been used, and confusion has often arisen as a result. This position should improve as the use of the B.S.I. Glossary (British Standards Institution 1958) becomes more widespread; this retains only one definition, that the speed of a pump for a given gas is the ratio of the throughput of that gas to the partial pressure produced by it at the mouth of the pump.

Apart from confusion of definitions, however, it has been difficult to compare the results of different workers because of the different methods used for experimentally measuring pumping speed. Many methods have been subject to inherent errors, and the fact that this has been recognized and the nature of the errors understood represents appreciable progress during the last ten years.

Attention was first drawn to the source and size of common discrepancies by Dayton (1948). He pointed out that a method of measurement used in many laboratories employed a flat plate to 'blank off' the mouth of the pump, the gas being admitted through a small hole in the plate. He showed that this method could result in high values of pumping speed being obtained due to the gas molecules being 'beamed' directly into the pump without the normal amount of diffusion and scattering. To avoid this effect it was necessary to use, instead of a plate, a test dome of diameter and length at least equal to the pump diameter, and to introduce the gas at the top in a non-directional manner. Dayton also investigated the effect of the direction of the opening in the tube leading to the

pressure measuring gauge, and showed that for a given throughput a lower pressure was recorded when the end of the gauge tube faced the mouth of the pump than when it faced the top of the dome, as would be expected from the relative numbers of molecules moving in the two directions. Figure 10 shows the magnitude of the speed difference for the different gauge tube directions as measured for a condensation pump of 4 in. diameter.

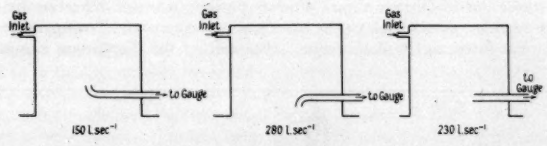


Figure 10. Variation of measured speed of a pump with direction of gauge tube opening. (*After* Dayton.)

More recently, Dayton *et al.* (1955) have made further recommendations on procedure for speed measurement and have criticized the method of Oatley (1954) on the grounds of the beaming effect of the variable holes employed, which although they are not in a plate immediately on top of the pump are separated from the pump mouth by a tube which "is sufficiently wide and short to have negligible effect on the speed of pumping".

A detailed discussion of the influence of gauge connection and position on the pressure values obtained in pump speed measurement has been given by Venema (1954). He is in agreement with Dayton on the nature of the effects found, but he proposes the adoption of a method giving a different result.

Although there is yet no agreed procedure for pump speed measurement, the present position is an advance on that of a few years back in that the possibilities of error are more widely recognized and the relationship between speeds measured by different methods is better understood. Provided that details of the measurement techniques are given when speed data are reported, the end of years of confusion and ambiguity is in sight.

§7. Measurement of Pressure

7.1. *Mechanical Gauges*

Gauges employing a diaphragm as a pressure-sensitive element in place of a column of liquid have been used for many years. More recently the need to remove the contamination of manometer fluid from systems handling pure gases, as well as requirements for robustness and remote indication have resulted in many improved designs of diaphragm gauge. Many designs employ the change of capacitance between the diaphragm and another electrode as a measurement of deflection and a theoretical investigation of the characteristics of a simple system of this type has been made by Lilly *et al.* (1947). The effects of ambient temperature changes have also been investigated theoretically and experimentally (Pressey 1953). Typical gauges are described by Cook and Danby (1953), Baxter (1953), Opstelten and Warmoltz (1955), and Opstelten *et al.* (1956).

The construction of the gauge used by the last-mentioned workers is shown in figure 11. The metal parts are of stainless steel with copper gaskets, and no rubber or plastic is used in the construction. The diaphragm D is of stainless steel, .20 microns in thickness, and is clamped, at a diameter of about 2·5 cm, between stainless-steel flanges F, the seal being formed by copper gaskets. The diaphragm forms a pair of differential condensers with discs C_1 and C_2, insulation being effected by annular ceramic rings I. The condensers form part of a high-frequency bridge circuit whose out-of-balance signal, after amplification and rectification by a phase-sensitive rectifier, is fed back to the manometer condensers to compensate electrostatically the force on the diaphragm. One side of the diaphragm communicates

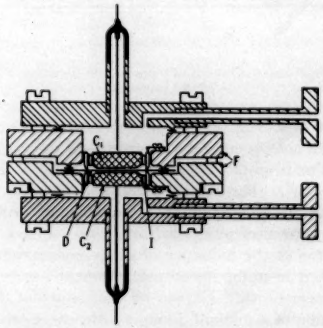


Figure 11. Diaphragm type pressure gauge. (*After* Opstelten *et al.*)

with the pressure to be measured, the other is evacuated, a low pressure being maintained by the 'cleaning-up' action of a Penning gauge. When used with the feedback circuit the claimed range is from 0·7 mb to about 2×10^{-5} mb. It can also be used without feedback by measuring the change of capacitance, when it is not as sensitive but can measure higher pressures up to 10 mb or higher, depending on the elastic constants of the diaphragm.

Alpert *et al.* (1951) have described a null type of diaphragm gauge where the diaphragm separates the pressure to be measured from a variable reference pressure measured by a normal oil manometer, the capacitance between the diaphragm and an electrode serving to indicate the position of equilibrium. A perforated backing plate limits the movement of the diaphragm, so that the reference side may be removed while the rest of the gauge is baked on the vacuum system.

Other gauges have used the variation of mutual conductance as an indication of diaphragm movement, and the advantages of linearity of calibration over a large pressure range and of high sensitivity have been claimed (Dibeler and Cordero 1951, Greenough and Williams 1951).

To sum up, the recent literature shows that diaphragm type gauges made of clean materials, bakeable, with robust electrical indication and good stability are

feasible. Scales from low pressures to several millibars are possible, with discrimination certainly to better than 10^{-3} mb. Such gauges have not been developed on a commercial scale, however, but have been designed individually to meet specialized needs.

7.2. *Acoustic Gauges*

There has been a recent suggestion for a gauge covering the 'mechanical' gauge pressure range but depending on the acoustic properties of the gas (Millar 1956). The gauge is designed so that the gas constitutes an acoustic transmission line, a signal introduced by an electromagnetic transmitter at one point being received by a similar receiver at another suitable point, say the far end of a tube. Since the attenuation depends on the gas pressure, the output is a measure of pressure, and it is claimed that the gauge can be designed to have a linear relationship over a wide range. Transmission along the tube itself must be minimized by the use of suitable materials or by more complicated design. Pulse operation, with a pulse short compared with the period of the transmitter and receiver diaphragms, is preferred.

Another recent paper describes the principle of an 'acoustical' gauge (Schwarz 1956) but the title here is somewhat misleading, as the proposed gauge is a refinement of the well-known decrement gauge. It is proposed to drive a membrane at its resonant frequency from a constant amplitude oscillator, the force being either electrostatic or electromagnetic, and by suitable circuitry to measure the amplitude of oscillation. A prototype calibration shows a measurable amplitude change from atmospheric pressure to below 10^{-3} mb.

Further results on this type of gauge and on that of Millar should prove interesting. In both cases the details available are as yet insufficient to allow any forecast of their future importance.

7.3. *Thermal Conductivity Gauges*

Progress in the Pirani type of thermal conductivity gauge can be said to amount to consolidation and completion of knowledge existing from its early days. Many workers have contributed to investigation of the effect of such variables as the cross section, material, length and working temperature of the filament, and of the material and size of its containing bulb. Perhaps it is not unjust, however, to single out the work of Von Ubisch for special mention, both for its completeness and its usefulness. The extension of the range upwards to the order of 100 mb for air is due largely to his work; the extension is made possible by using a feedback circuit to adjust the power input and keep the wire temperature constant. (When input conditions to the bridge are kept constant the pressure range is limited because the wire approaches room temperature at a pressure of the order of 1 mb.)

A summary of the work giving the important results and references to more detailed publications has been given by Von Ubisch (1952).

Benson (1956) has pointed out that the thermocouple type gauge, unlike the Pirani, normally includes within a single bulb means for compensation for steady-state ambient temperature changes. The gauges described by him have the further advantage of compensating for transient effects of ambient temperature

changes. Their elements consist of alternate sections of the unlike metals butt-welded together, and there are no separate heating elements, heating being effected by passing alternating current through the couple wires. A number of couples are connected in an array that forms part of a bridge circuit, balanced for a.c. but able to indicate the d.c. output from the couples. Transient compensation is achieved by heating some of the couples and leaving others unheated, the circuit being arranged so that the outputs of the heated and unheated couples are in opposite polarity. A sudden change in ambient temperature produces transient e.m.f.s in all the couples, but those produced in the unheated couples compensate those produced in the heated ones. Multiples of three couples are used, as many as fifteen having been employed for special applications. The pressure range covered in a given bulb is varied by choice of wire size and heating current. Two of the designs cited by Benson cover between them the range from 30 mb to about 10^{-3} mb.

7.4. *Knudsen Type Gauges*

In the fifty years since Knudsen's work many designs of Knudsen type gauges have been published, and descriptions of representative constructions are given in the textbooks (Dushman 1949, Leck 1957). Recent advances have been aimed at reducing the delicacy and susceptibility to vibration which in most designs is the penalty incurred for high sensitivity. High sensitivity, requiring low restoring force on the moving element or vane, also results in a long period of oscillation. The addition of a feedback system of control which maintains the moving element very nearly in its zero position, measuring gas pressure in terms of the control force required, eliminates these disadvantages. The use of such a feedback control system brings further advantages; the pressure range can be increased, since it is not dependent on mechanical limitation of movement of the moving element; a much stronger suspension system can be used, since its restoring torque will be very small due to the negligible deflection allowed; speed of response can be further improved by controlling damping with suitable control circuits; indication by spots of light or microscope observation is no longer necessary, and remote indication becomes possible.

A gauge and control system having these advantages has been described by Litting and Taylor (1952) and their monograph will be of great help to designers of control systems of this type.

In essentials, their moving element consists of two vertical aluminium vanes at the ends of a light rod which is suspended by a taut tungsten wire. The vanes are each situated between a pair of aluminium plates, made of foil stretched over ceramic rods. One pair of plates is used to provide the pressure-dependent force, the other the balancing electrostatic force.

The two capacitances between the 'control' vane and its plates are compared by an a.c. bridge the output of which is proportional to the error between the actual position of the vane and its central position when the bridge is balanced. This error voltage is amplified, rectified and applied back, after stabilization, to the capacitor plates to produce an electrostatic force opposing the deflection. If the amplification of the error signal is sufficient there is no appreciable rotation of the vanes and the direct voltage applied to the plates is proportional to the pressure-dependent force, and can be used as a pressure indication.

During calibration the lowest pressure reached by Litting and Taylor was about 7×10^{-7} mb, corresponding to a capacitance variation of 5×10^{-11} microfarad, but the smallest detectable change was fifty times smaller than this, so that a pressure approaching 10^{-8} mb would be measurable. A linear relationship, independent of nature of gas, between pressure and output voltage was demonstrated up to pressures of the order of 5×10^{-2} mb, and the gauge was usable to higher pressures if the nature of the gas was known.

A control system of this type, applied to a smaller, easily degassed gauge and made available commercially would fill a very long-standing need.

7.5. *Radioactive Ionization Gauges*

An ionization gauge in which alpha particles from a radium source formed the ionizing agency was first described by Downing and Mellen (1946) and modifications for special purposes were mentioned later by Mellen (1948).

The early gauge, which has found considerable industrial use, under the name 'Alphatron', had as an alpha source a piece of a gold radium alloy containing about 0·2 mg of radium, in a cylindrical ionization chamber about 3 in. long and 3 in. in diameter, giving a sensitivity of about 2×10^{-10} amp per millibar for air. The advantages are linearity of calibration, freedom from the clean-up and decomposition effects of a hot filament and freedom from damage due to inadvertent admission of air. The original 'Alphatron' gauge had three pressure ranges of 10, 1, and 0·1 torr full scale. These ranges have been extended in later models, to a maximum above atmospheric pressure and to an improved sensitivity lowest range of 0·01 torr full scale, the improvement being due partly to the use of two ionization chambers of different sizes. This and other recent advances in this type of gauge and its associated circuits have been described by Vacca (1956), who discusses also the factors influencing linearity and speed of response. Industrial users of the earlier gauges will appreciate the mechanical improvements in the new design, where the ion collector is better supported and the gauge consequently much less affected by shock and vibration.

7.6. *Measurement of Very Low Pressures*

An area of 1 cm^2 exposed to a pressure of 10^{-7} mb of air at room temperature is struck by about 3×10^{13} molecules per second. For oxygen molecules a complete monolayer is formed by about 8×10^{14} molecules per cm^2, so that at an air pressure of 10^{-7} mb a surface with a condensation coefficient of unity for oxygen will be covered with a monolayer in less than two minutes. Before 1940 such workers as Langmuir, Anderson and Nottingham had obtained pressures so low that the drift of work function of a freshly prepared emitting surface was small in a time of the order of an hour (Dushman 1949). It was noticed by these early workers that such drift times were better indicators of the pressure in their apparatus than were the readings of their ionization gauges, and that however low the pressure estimated from surface contamination considerations, the ionization gauge never indicated a pressure as low as 10^{-8} mb. By 1947 the realization was growing that there was a physical limit to the lowest reading of an ionization gauge, and Nottingham proposed an explanation, later shown to be correct. His hypothesis was that photo-emission from the ion collector is the cause of a current

indistinguishable from that due to positive ions arriving at the collector; the impact of electrons with the accelerating electrode gives rise to X-rays which are the main cause of the photo-emission from the ion collector.

Conventional ionization gauges were nevertheless employed to measure lower pressures by the use of the 'flash filament' technique (Apker 1948, Hagstrum 1953). The ionization gauge is provided with an additional tungsten filament, which is first cleaned by flashing to a sufficiently high temperature. After a measured time at room temperature the filament is again flashed, releasing the molecules adsorbed on it and giving a pulse pressure reading on the ionization gauge. This pressure pulse increases with time between flashes up to a saturation

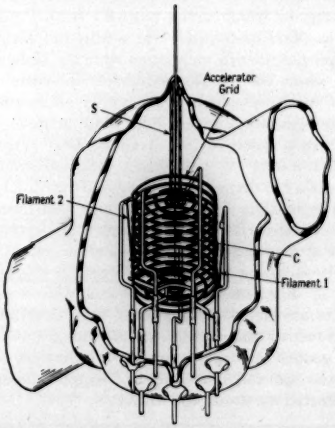


Figure 12. A design of Bayard–Alpert type ionization gauge.

value above which increase in time produces no increase in pulse pressure. The minimum time interval to give this saturation state is assumed to be the time taken to form a monolayer, and the pressure is calculated accordingly. This technique has proved very useful, but it has the disadvantage of indicating only the partial pressure of those gases whose condensation coefficients are high. Also it requires a considerable time for each pressure measurement.

Appreciation of the reason for the low-pressure limit of the conventional ionization gauge quickly led, however, to designs which substantially reduced the photo-emission effect (Bayard and Alpert 1950, Lander 1950, Metson 1951). The Bayard–Alpert type of gauge, which has been used most widely, employs a fine wire as ion collector, thus reducing the solid angle presented to X-rays from the grid by a large factor (about one thousand times smaller than in a conventional gauge with cylindrical collector). To retain sensitivity the collector is placed centrally inside the accelerator grid and the filament is outside. The collector is

slightly negative (say -25 volts) and the accelerator about 150 volts positive. The collector is thus in a potential well, and ion collecting efficiency is as high as in a conventionally constructed ion gauge. A typical design of Bayard–Alpert gauge is shown in figure 12.

The ion collector C is made from 0·15 mm diameter tungsten, tapered from its base by a deplating process, to minimize the solid angle presented to and to reduce any tendency to mechanical oscillations. The glass tube shield S round the collector lead-in prevents x-rays impinging on this larger diameter wire, lengthens the surface electrical leakage path to the collector and prevents leakage due to any evaporated metal deposited on the walls.

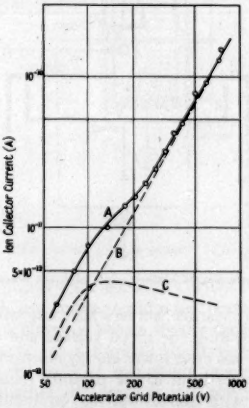


Figure 13. Components of ion current in Bayard–Alpert gauge. (*After* Alpert and Buritz.)

Bayard and Alpert determined the magnitude of the x-ray effect contribution to the apparent ion current in their gauge by plotting the apparent ion current I_c against the accelerating voltage E_a. When the x-ray effect is predominant the curve of $\log I_c$ against $\log E_a$ was found to be a straight line of slope approximately 1·8; when the true ion current was large compared with the photo-emission the curve was very similar to the curve of probability of ionization, rising to a maximum between 100 and 200 volts and falling slowly with increase of potential. The continuous curve A of figure 13 gives the total current as a function of voltage, and corresponds to an indicated pressure of about $1{\cdot}3 \times 10^{-10}$ mb. The broken curves show how it is made up from the straight-line x-ray effect (curve B) and the ionization current (curve C). At 150 volts the current due to x-ray effect is approximately twice that due to ionization, so that the true pressure is of the order of 4×10^{-11} mb.

Figure 14 illustrates diagrammatically the principle used in Metson's ionization gauge to suppress the photo-emission. The gauge is a tetrode with cathode C,

accelerator A, cylindrical ion collector K and a suppressor S, also cylindrical and coaxial with K. The suppressor is provided with a slit, situated opposite the cathode. The collector is at a potential of -2 volts, the accelerator at 200 volts. Ions and x-rays both reach the collector through the slit in the suppressor but the latter, at -40 volts, provides a retarding field preventing any photo-electrons from leaving the collector. Metson showed that a retarding field of 30 volts was sufficient to suppress all photo-electrons within the limits of sensitivity of his experiment. This maximum photo-electron energy is less than would be expected

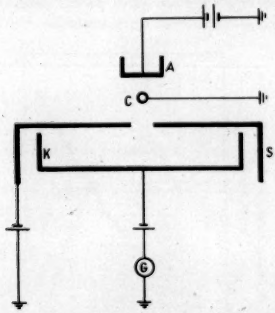


Figure 14. Principle of Metson's low pressure ionization gauge.

from the Einstein equation ; earlier workers have suggested that this effect is due to the x-ray photons penetrating the metal surface and transferring their energy to electrons, each of which can eject lower energy electrons from the metal surface. There appears to be no lower limit to the pressure measurable with the Metson gauge, as there must be with gauges relying on reduction of the solid angle presented by the collector. There would be a limit if the x-rays could reach the collector side of the suppressor by reflection from the collector, and Metson searched for evidence of such reflection but did not find any.

The latest work on gauges for very low pressures, reported by Redhead (1957) and further by Hobson and Redhead (1958) is on a modification of the Penning type cold-cathode ionization gauge. Many designs of Penning gauge in use fail to strike at pressures much below 10^{-5} mb, but that this is not a fundamental limitation was shown as early as 1942 by Klemperer and more recently by the design of Beck and Brisbane (1952). Hobson and Redhead point out that there is no theoretical x-ray limit in a Penning gauge since the electron current producing the x-radiation is proportional to pressure, but that in normal designs there is another limitation : the cathode is used not only as an ion collector but also provides the initial field emission to start the discharge, and this field emission is indistinguishable in the measuring circuit from the ion current. In the Redhead design the two functions are separated by the provision of additional cathode electrodes which provide the starting field emission and shield the ion-collecting cathode from high electric fields. The construction by which this is achieved is

shown in figure 15. The gauge is of the inverted magnetron type, with an anode of 1·0 mm tungsten wire situated along the axis of the ion-collecting cathode, which is a cylinder of Nichrome V 3·0 cm in diameter and 1·35 cm long, with closed ends, through holes in which shield tubes pass. These shields, whose ends are electropolished to reduce field emission, are in contact with the auxiliary cathode, which is a box surrounding the whole assembly. The gauge is operated with an anode potential of 6 kv and a magnetic field of about 2000 oersteds.

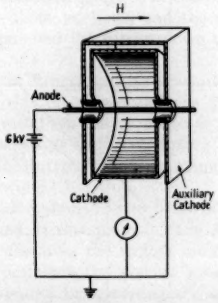


Figure 15. Section of inverted-magnetron gauge. (*After* Hobson and Redhead.)

Over a wide range of pressure the collector current obeys the relation $I_c = Kp^n$ where p is the pressure, K is a constant and n lies between 1·1 and 1·4 for various gauges. The sensitivity achieved is shown by the calibration curves of figure 16, which also shows the curve for a Bayard–Alpert type of gauge. It is evident that a sensitivity at least an order of magnitude greater than the Bayard–Alpert gauge is attainable. A further advantage is that there is no lower limit due to the vapour pressure of a hot filament (Alpert and Buritz 1954).

As this is being written a further design of very low pressure Penning type gauge has been announced by Redhead (1958). This gauge, having an outer cylindrical anode, is termed by Redhead the magnetron gauge to distinguish it from the inverted magnetron type. Its electrode construction is shown in the cut-away sketch of figure 17. The anode cylinder, 1·4 cm long and 3·0 cm in diameter, is perforated to improve access of gas to the gauge. The cathode is shaped like a spool, with an axial cylinder 2 cm long and 3 mm in diameter, welded to two end discs. Two annular auxiliary cathodes shield the cathode end discs from the electric field at the ends of the anode cylinder. Their edges are rounded and electropolished to reduce field emission. The glass envelope of the gauge has a conducting film deposited on its inner surface. This forms an electrostatic shield for the cathode discs and is connected to the auxiliary cathodes by a spring contact.

The gauge is operated at an anode potential of 5 or 6 kv in a field of 1000 oersteds, and has been used down to pressures of 10^{-13} mb (compared with 3×10^{-12} mb for the inverted type). It has a higher sensitivity than the inverted type, a figure

of fifty times that of a Bayard–Alpert gauge being quoted. Over part of the pressure range, from the highest pressures to about 5×10^{-10} mb the relation between collector current and pressure is linear, but below that pressure it suddenly changes and follows the relation $I_c = Cp^{1.7}$. The reason for this is not clear, and experiments are proposed to establish whether the change of slope may be preventable by choice of suitable operating parameters.

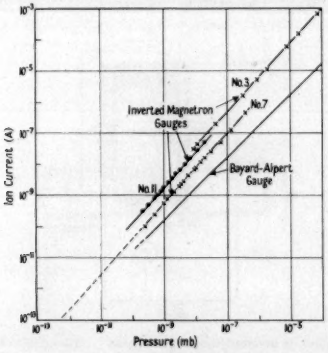


Figure 16. Sensitivity calibrations of inverted-magnetron gauges. (*After* Hobson and Redhead.)

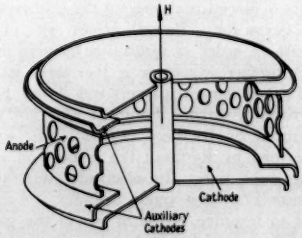


Figure 17. Magnetron type low pressure gauge. (*After* Redhead.)

There is no definite indication of the lower limit of pressure measurable by these gauges ; it could well be set by noise in the collector circuit.

§ 8. Very Low Pressure Techniques

The developments in gauges which enable very low pressures to be measured having been discussed in the previous section, it is appropriate to review briefly the techniques used to attain such low pressures. Their attainment by earlier workers using sealed-off techniques has already been mentioned, but the steps

forward which have led to their routine use in laboratory vacuum systems began with the work of Alpert (1953). A glass system using a fractionating oil condensation pump was employed, the working part of the system containing a Bayard–Alpert gauge. This part of the system could be isolated from the oil pump by means of a bakeable metal tap (previously described by Alpert 1951). The procedure was to pump the system to a pressure of the order of 10^{-7} mb with the oil pump, and bake the working part and the metal tap for several hours (usually overnight). The ionization gauge was then degassed at electrode temperatures of 1200°c, taking about an hour, and the metal tap closed. Normal running of the gauge then pumped the system down to pressures of the order of 10^{-10} mb within an hour.

Using systems of this sort, Alpert and his co-workers were able to carry out much useful work in the fields of surface and gas discharge physics. For the admission of pure gases the metal tap was used also as a variable leak, and enabled gas purities to be maintained which had never before been conveniently possible. High gas pressures were measured without contamination with the null type diaphragm gauge referred to in §7.1.

Results obtained with such systems in two fields of interest to vacuum physics should perhaps be mentioned at this point. They have been given by Alpert and Buritz (1954) in a discussion of the factors limiting the attainment of very low pressures. In one experiment the vapour pressure of tungsten between 2600°K and 3300°K was measured by operating the spare filament of a Bayard–Alpert gauge at temperatures in this range while keeping it at a positive potential to prevent emission of electrons. The gauge was used in the normal way with the other filament to measure the equivalent nitrogen pressure of the tungsten atoms. The effects of heating of the envelope were avoided by immersing the gauge in liquid nitrogen. The results when plotted gave a value for the heat of vaporization of tungsten agreeing within 0·7 % with Reimann's (1938) value. The equivalent nitrogen reading of about 10^{-12} mb at filament operating temperature agreed well with an estimate made assuming a value for the ionization probability of tungsten equal to that of mercury.

The other experimental work concerned the diffusion of helium through glass, following the detection of a constant rate of rise of pressure of inert gas in very low pressure systems, independent of degassing procedures. The enclosure of a gauge in an external vacuum jacket stopped the rise of pressure after about 60 days, proving the external origin of the gas, and further experiments resulted in values of diffusion coefficient and solubility in agreement with results available for helium, but covering a wider range, since the ability to measure very small pressure changes enabled permeation rates to be measured at much lower temperatures than had previously been possible (thus eliminating, incidentally, some previous sources of error). The value for the diffusion coefficient of helium in Pyrex glass at 27°c was found to be 10^{-8} cm^2 sec^{-1}. Independent check that the gas was in fact helium was obtained by use of an omegatron (see §9).

Alpert's original method for obtaining low pressures by isolating the working part of the system with its ion gauge pump from the vapour pumping system was in effect, as far as an individual experiment was concerned, a sealed-off system, and the work which could be carried out in such systems was limited by the

pumping speed of the gauge (about 1 litre sec^{-1} for the active gases and 0·01 to 0·1 litre sec^{-1} for others).

This limitation was imposed by the need to exclude contamination coming from the oil pump, and it was later found possible to dispense with the isolating tap (Alpert 1954) and replace it by the copper foil trap referred to in § 4. It was shown that the oil condensation pump continued to pump at pressures below 10^{-10} mb, and pumping speeds for inert gases were therefore increased (but were limited by the conductance of the system to about 0·1 litre sec^{-1}). Redhead has used a glass mercury vapour pumped system with liquid nitrogen trapping to achieve similar pressures. These results raise the possibility of reaching very low pressures with systems having much higher speeds, and thereby increasing the scope of the work possible. The author and D. S. Stark have experimented with systems using trapped mercury vapour pumps in both glass and stainless steel. In a stainless-steel system with demountable copper gaskets, pumped by a single-stage stainless-steel mercury vapour pump, pressures of 10^{-10} mb have been obtained. To achieve sufficient exclusion of mercury vapour it was found necessary to use two of the conduction type traps (of § 4) in series. The whole of the system down to the top of the pump was first baked and the oven then partly raised so that the lower trap could have coolant applied while the rest of the system continued to bake. Finally the oven was raised completely and the second trap cooled. The size of traps used was such that the overall pumping speed obtainable was about 10 litres sec^{-1} down to pressures of 10^{-6} mb, but measurements of the speed at the low pressures have not been made. The experiments illustrate, however, that the very high vacuum field is not limited by glass systems and low conductance traps.

The advent of the Penning type ionization–gettering pump described by Hall (see § 5.1) brings another possibility of increased pumping speed at very low pressures without the need for traps or restrictions, although there are at the time of writing no published data on its low pressure limitations.

The lowest pressures so far used have been obtained by the use of liquid helium trapping. A portion of a very low pressure system cooled with liquid helium pumps all gases incident on it, including helium itself. A system of this sort has been described by Hobson and Redhead (1957), who showed that the equilibrium pressure above approximately 10^{-6} monolayer of helium adsorbed on glass at a temperature of 4·2°K was less than 10^{-12} mb.

§ 9. Newer Mass Spectrometers for Vacuum Systems

An account of progress in vacuum technique would not be complete without some reference to new types of mass spectrometer suitable for investigating gas composition in vacuum systems. The use of conventional deflection types in vacuum work is well known, and will not be discussed here. Space will permit only an outline of the principles of some other types which can be attached to a vacuum system as a measuring tool, almost as easily as a pressure measuring gauge can be attached.

9.1. *The Omegatron*

Since the omegatron was first used for accurate measurement of the charge–mass ratio of ions (Hipple *et al.* 1949, Sommer *et al.* 1951) it has been applied,

in simplified form, to a variety of tasks in vacuum systems (Morgan *et al.* 1953, Alpert and Buritz 1954). In mode of operation it bears resemblance to the cyclotron. Ions are formed along the path of a fine electron beam, which is parallel to a magnetic field H, and are accelerated by an alternating electric field between two parallel plates which correspond to the dees of a cyclotron and are known as r.f. plates. When the applied frequency is equal to the cyclotron frequency, $f = He/2\pi m$, the ions of charge e and mass m gain energy at each revolution and their orbits increase in radius until they can be collected on a suitably placed

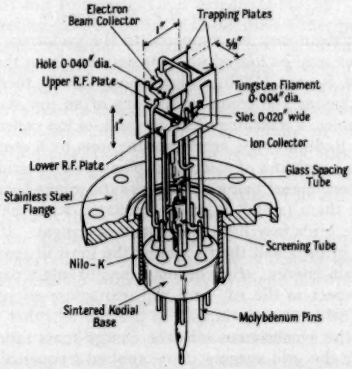


Figure 18. Construction of a simple omegatron.

electrode, the current to which is a measure of the abundance of ions of the charge–mass ratio selected by the value of applied frequency. Berry (1954) has shown that ions of a given ratio are bunched in space regardless of their starting phase.

Details of the construction of a simple omegatron similar to that described by Bell are shown in figure 18. The ends of the r.f. plates are folded over to improve the uniformity of their field, and the interaction chamber is completed by two end plates, having holes for entry and exit of the electron beam. These plates are biased to prevent loss of ions in directions parallel to the beam and are termed trapping plates. The ion collecting electrode is a thin strip parallel to the beam near one r.f. plate, the lead to it being shielded both inside and outside the bulb to minimize electrical pick-up in the current measuring circuit.

The convenience and small size is evident from the diagram, and for many purposes a large magnet is not required ; a field of 2400 oersteds is typical. The sensitivity is of the same order as that of a normal ionization gauge, but beam current being only a few microamps means that a vibrating-reed electrometer is necessary to measure the collector currents. The resolution varies as $H^2er/2mE$,

according to the simple theory, and for a given instrument can be appreciably improved only by increase of H, as decrease of r.f. field E produces loss of sensitivity. Even with the low field of 2400 oersteds, however, the argon and carbon dioxide peaks can be resolved. For higher masses resolution falls off as predicted, but at fields of the order of 5000 oersteds, adjacent masses are separated from the CO_2 peak and the instrument is usable up to mass 200.

The simple omegatron finds its main use in systems operating at pressures below 10^{-5} mb, and preferably below 10^{-6} mb, since background current and inconstancy of characteristics become troublesome with increased pressure.

9.2. *The Linear R.F. Spectrometer*

Just as the omegatron may be compared to the cyclotron, another small type of mass spectrometer may be looked upon as deriving from the linear accelerator. Since the papers of Bennet (1950) and Redhead (1952) there have been many variants described, having the common features of an ion source, followed by a system of r.f. electrodes, a retarding electrode and an ion collector. Those which are basically of the Redhead type accelerate the ions by a steady potential into a r.f. electrode system consisting of a series of plane parallel grids. Alternate grids are connected together, one set being earthed and the others having a r.f. alternating potential applied to them (which is small compared with the initial accelerating potential, so that the grids may have equidistant spacings). If the transit time of an ion between two grids is half the r.f. period the ion will remain in synchronism with the field and gain energy. Synchronous ions having a particular phase angle at entrance with respect to the r.f. field will gain more energy than any others, so that a retarding field can be arranged to prevent all other ions from reaching the collector. For the synchronous ions the charge–mass ratio is proportional to S^2f^2/V_0, where S is the grid spacing, f the applied frequency and V_0 the steady accelerating potential. It can be shown that ions travelling at sub-multiples of the synchronous velocity are rejected if it is arranged that the time such an ion spends in the r.f. field is an appreciable part of a cycle.

Sensitivity depends on the ionizing efficiency of the ion source and upon the overall transparency figure for the grids, so that wide variations between different designs are to be expected. Bennet's grids, for instance, were made of 0·012 mm tungsten wire at a mesh of 15 per cm, and thirteen of them passed more than 50% of the entering beam. Since the retarding potential is normally set to exclude all ions except those within a certain range of phase, however, an overall transmission efficiency of 10% of ions formed is a more representative figure (Redhead and Crowell 1953, Boyd and Morris 1955).

Resolution is dependent on the number of stages and on the retarding conditions and r.f. input. For the Redhead type of N stages resolution is proportional to $N(N^2-1)$. Instruments have been described of all sizes from two stages to fifteen or more ; it is fair to generalize by saying that an instrument of this type with a resolution as good as the simple omegatron can be made with a linear dimension such that it is conveniently usable on a laboratory vacuum system.

The fact that no magnetic field is needed is of great importance, and is often the factor leading to the choice of one of these instruments or one of the related time-of-flight spectrometers.

9.3. *The Time-of-Flight Mass Spectrometer*

If ions from a low energy source are accelerated through a potential V they will reach a velocity given by $v = (2eV/m)^{1/2}$. If they are then allowed to pass through a field-free drift space of length L they will arrive at a collecting electrode at the end of the drift space after time intervals proportional to $m^{1/2}$, and the time separation between two ions of masses m_1 and m_2 will be proportional to $L(m_1^{1/2} - m_2^{1/2})$. Measurement of the time by means of cathode-ray tube and time-base, or by gating pulse techniques completes the spectrometer. This type has found wide use, typical references being Cameron and Eggers (1948), Wolff and Stephens (1953) and Katzenstein and Friedland (1955). Wolff and Stephens pulsed their accelerating field to obtain a linear mass scale, pulse lengths of from 5 to 50 microseconds covering the mass ranges required. Katzenstein and Friedland's instrument was a large one, occupying a tube one metre long and 6½ cm in diameter, and resolving adjacent masses above mass 100. Wiley and McLaren (1955) achieved this order of resolution with a drift space only 40 cm long, by using a double field ion source. Theoretically, of course, the gated time interval allowed for ion collection around the end of the time-of-flight of a given mass may be made as small as is necessary for any required resolution, and the limit is then set by the spread of thermal velocities of ions in the ion source.

For many applications there is little to choose between the time-of-flight mass spectrometer and the linear r.f. instrument. Where a rapidly changing mass spectrum is to be observed or recorded the ease of display and speed of presentation of information possible with the time-of-flight instrument without special measuring techniques is an advantage.

§10. Leak Detection

Two decades ago there would have been little to discuss under the heading of leak detection. Leaks in glass systems were found with the Tesla coil, but the experimenter finding small leaks in a metal system was almost helpless. The common method of detection by change in colour of the discharge in an attached glass tube when the leak was covered with some volatile liquid such as benzene was insensitive and very time-consuming. As vacuum systems became larger and more complicated, and as the requirements of the atomic energy programme grew, the inadequacy of the existing leak-finding methods became obvious and investigation of improved methods began. Subsequent work has resulted in many ways of finding leaks; modern requirements for large and expensive sealed-off apparatus of metal construction, such as high power microwave valves, have created a need to detect even smaller leaks than those of importance in pumped systems, and continuing work to improve detection sensitivity has been stimulated.

Reviews of leak detection practice have been given by Jacobs and Zuhr (1947), Dushman (1949) and Blears and Leck (1951 a). The use of liquids in leak finding has now been given up almost completely in favour of the use of gases, employed in fine jets for searching, or under hoods or bags to determine whether leaks exist in a piece of apparatus. Liquids have numerous disadvantages—they may block a leak completely and in the case of rubber gaskets can cause swelling and so mask a leak, while they increase response time by the time taken to evaporate from the leak.

10.1. *Detectors using Ordinary Vacuum Gauges*

Some vacuum gauges respond differently to different gases, and so are usable as leak detectors if the air passing through a leak can be replaced by a suitable gas. Thermal conductivity type gauges, for instance, have a higher sensitivity for hydrogen, and the effect is enhanced by the increased conductance of the leak for hydrogen. Ionization gauges can be used similarly with search gases such as butane. Methods of this type are usually made more sensitive by backing off the indicated current corresponding to the steady pressure and using a more sensitive indicating instrument to display the change ; a limit is set by random fluctuations of pressure in the system, and this has led to the development of differential methods of leak detection. In these methods, which have been reviewed by Blears and Leck (1951 b) identical gauges are connected to the same point in the vacuum system and their connections arranged so that both the mean pressure and the pressure fluctuations cancel out. The search gas is permitted to enter one gauge and excluded from the other so that a differential change of pressure is indicated when a leak is found. This method has been used with thermal conductivity gauges and with both hot and cold cathode ionization gauges, and various means of excluding the search gas from one gauge have been employed. Butane as a search gas excluded by liquid air trap and carbon dioxide excluded by a trap containing calcium hydroxide are typical examples.

10.2. *Detectors employing Special Properties of the Search Gas*

Other detectors do not depend on the change in an apparent pressure reading but detect the presence of the search gas by some other effect. The halide detector (White and Hickey 1948) is one example, employing the effect of a gaseous halide upon the positive ion emission of platinum. As this effect occurs in air the method can be used with the detector outside the system and the search gas under pressure inside. Another useful detector employs the poisoning effect of oxygen on the emission of a tungsten filament, and has the advantage of being usable without much extra apparatus in many cases, as the emission of the normal ionization gauge fitted to the vacuum system can be backed off and changes displayed on a sensitive meter. Oxygen is a readily available search gas, safe and convenient. This method is at least as sensitive as the palladium barrier ionization gauge method, which also falls in this general class of detectors. Here the search gas used is hydrogen, and advantage is taken of its rapid permeation through heated palladium. A well evacuated and gettered ionization gauge is separated from the leak detecting system by a palladium barrier, which can be heated to 700–800°c, when hydrogen can pass through and provide an indication independent of the air pressure in the system and of pressure fluctuations. A disadvantage is that oxygen adsorbed on the palladium can slow down the permeation rate of hydrogen unless the correct proportion of oxygen to hydrogen is ensured by introducing an auxiliary leak of one of the gases ; on the credit side is the fact that it is very suitable for use with the 'backing space technique' where the detecting element is placed on the backing side of the vapour pump to take advantage of the compression effect of the pump in multiplying the pressure of the search gas—an effect which can be large if the volume containing the detector is small and can be isolated from the backing pump.

10.3. *Mass Spectrometer Methods*

Mass spectrometers are an obvious choice as leak-detecting instruments; they can be set up to respond only to the search gas, and allow the choice of a search gas which is not normally present in the system, which will pass rapidly through a leak and which will not contaminate the system in any way. Helium is the usual choice when all these advantages are required. Conventional deflection type instruments have been widely used and are commercially available. Both 60° and 180° deflection instruments have been employed, typical early designs having been described by Worcester and Doughty (1946) and by Thomas *et al.* (1946). An interesting 180° type employing a cold-cathode ion source, and tuned a.c. amplification made possible by modulation of the ion beam velocity, has been described by Milner (1951). The simplicity and portability which has been achieved with deflection-type instruments after ten years of development is well illustrated by Charpentier (1956), who gives an account of a 180° type which is completely self-contained, with its pumping system, in a cabinet measuring 19 in. × 22 in. × 20 in. With the use of only six valves in its amplifier circuitry and power supplies the sensitivity is sufficient to detect the helium in normal air (about 4 parts per million).

The newer types of mass spectrometer described in § 9 have also been applied very successfully to leak detection. A small and simple linear r.f. spectrometer tube for leak detection has been used by Varadi and Sebastyen (1956) and a more elaborate example of this type which is commercially available has been reported on by Moody (1956).

The omegatron also has proved to be a sensitive and convenient detector of small leaks, and has been discussed in this context by Bell (1956), though not many workers would agree with the reasoning leading to his claims that the omegatron has a large advantage over other mass spectrometers as a leak detector.

10.4. *Performance of Leak Detecting Systems*

It will have been noticed that in the previous sections no figures of sensitivity have been quoted, although most publications on leak detectors make claims for the size of leak which can be detected. These figures have purposely been omitted because there is at present no satisfactory way of comparing different detecting systems. A detecting system comprises, in addition to the component under test, a pumping system, a leak-detecting element, with its ancillary equipment such as power supplies and indicator, and the search gas with the means of applying it. All the parts affect the efficiency of the system and even for a given detecting element and its ancillary equipment the smallest leak which can be detected will vary with the position of the element in the system, with the pumping speed and, for some elements, with the state of cleanliness of the system and the amount of other leakage present. Reducing the effective pumping speed by throttling will increase sensitivity by allowing a greater build-up of search gas, but the rate at which this is removed when the leak is no longer covered will be increased, and a limit is set when the time of response becomes too long for convenient working. The criterion of acceptable response time has varied from worker to worker and is often not stated at all in the published work. Nor is it usually discernible how near to the ultimate noise limit the indicating circuits of

the detector are operating, while the ratio of signal to background acceptable for detection is also not agreed. It is therefore almost an impossibility to give any sort of quantitative comparison of different types of leak detectors based upon the present published data. In general, it can be said, as some indication of the present position, that in a properly designed apparatus leaks of 10^{-10} litre mb sec^{-1} can be located, and, by using a long build-up time, the presence of leaks as small as 10^{-14} litre mb sec^{-1} can be verified. These results are obtainable with the better mass spectrometers, of all types, and probably with differential ion gauges pushed to their limit.

The unsatisfactory state of definition of leak-detecting performance has been recognized by the B.S.I. Sub-Committee on Terms and Definitions used in High Vacuum Technology, and the recommendations of B.S. 2951, Appendix A (British Standards Institution 1958) on this subject are worth noting as the first step forward in making comparisons possible. It is pointed out that because of the effects of the different parts of a leak-detecting system the definitions required should be divided, so that an assessment can be given of the sensitivity of a leak-detecting element separately from the performance of a complete leak-detection system. Thus there are two sensitivity definitions. The sensitivity of a leak-detection system is defined as the deflection of the indicator per unit leak completely covered with search gas (not more than three seconds being allowed between covering the leak and reading the deflection). The sensitivity of a leak-detecting element is the deflection of the indicator when, at the position of the element, unit partial pressure of air is exchanged for unit partial pressure of search gas (not more than three seconds to elapse between exchange of gases and reading of deflection). There are two further quantities which require to be known : the smallest leak detectable by a leak-detecting system and the smallest partial pressure exchange detectable by a leak-detecting element. These are not as easily defined, and for generality require to be looked upon as the sum of two quantities. The minimum leak detectable by a leak-detection system is defined as the sum of two quantities : (i) a quantity L_0, characteristic of the system when clean and leak-free, plus (ii) a quantity kL where k is constant and L is the total throughput present at the time of tests.

The minimum pressure exchange detectable by a leak-detecting element is also defined as the sum of two quantities : (i) a quantity p_0 characteristic of the element in a clean and leak-free system together with (ii) a further quantity Kp where K is constant and p is the mean air pressure when the change of air to search gas is carried out.

For all the definitions it is necessary to specify what signal level shall be taken as detectable, and it is suggested by B.S. 2951 that the signal shall be said to be detected if the deflection due to it is three times the mean peak-to-peak fluctuation, averaged over ten consecutive fluctuations.

Leak detection has progressed by being dragged in the wake of a rapidly expanding vacuum technology ; the adoption of some such definitions of performance as those quoted is now becoming essential for that proper comparison of methods which can stimulate further improvements in leak-finding technique.

§11. Conclusion

A review of progress must of necessity concern itself with the past ; worn by the technologist the mantle of the prophet fits but ill. Yet in the three basic aims of vacuum technology—to obtain a low pressure, to measure it and to maintain it during an operation—it is perhaps legitimate to cast a glance forward in the direction of our advance.

In producing our low pressure we are at the threshold of a new era. The vapour pump, which has made such great practical progress in the past thirty years and is still pre-eminent, is being challenged by newer methods. While it would be rash at this stage to forecast its complete disappearance, its inability to hold its position in a number of applications of modern vacuum technology must become apparent within the next decade.

Turning to measurement, the problem of measuring the very low pressures now obtainable has been solved, and the problem of knowing what gases are responsible for the pressure measured is not likely to remain a serious obstacle. The accuracy of our pressure measurements, however, even in normal apparatus, shows little sign of improving. In this field the problems need not be sought—they seek us out daily, and when we look for assistance we are not encouraged by the discovery that to no standards laboratory in the world can even a simple ionization gauge be sent for a certificate of calibration.

To maintain our low pressure in pumped, unsealed systems we need mainly sufficient pumping speed, and this will not be lacking. In sealed-off devices we shall be able to incorporate some of the newer pumping means to deal with rates of gas evolution which would have been far too great for the conventional getter. In both cases we shall be helped by the continuation of the trend to 'vacuum cleanliness' which is already apparent. Wax and grease have given way to synthetic elastomers, and these in turn are being replaced by metal seals and gaskets in great variety. The older methods of construction are yielding place to such techniques as argon arc welding of stainless steel. Baking of apparatus is being facilitated by the elimination of solders containing volatile metals.

Perhaps most encouraging is the new attitude of the vacuum technologist. Vacuum work has been beset by superstitions, unproved opinions and practices based on inadequate data. These are being swept away, and more attention to accurate measurements should hasten their going. Vacuum technology as a science and a tool will be the richer for their departure.

References

Alexander, P., 1946, *J. Sci. Instrum.*, **23**, 11 ; 1948, *Ibid.*, **25**, 313.

Alexeff, I., and Peterson, E. C., 1955, *Trans. National Symposium on Vacuum Technology*, 87 (London : Pergamon Press).

Alpert, D., 1951, *Rev. Sci. Instrum.*, **22**, 536 ; 1953, *Ibid.*, **24**, 1004 ; 1954, *Trans. National Symposium on Vacuum Technology*, 69.

Alpert, D., and Buritz, R. S., 1954, *J. Appl. Phys.*, **25**, 202.

Alpert, D., *et al.*, 1951, *Rev. Sci. Instrum.*, **22**, 370.

Anderson, P. A., 1937, *Rev. Sci. Instrum.*, **8**, 493.

Apker, L. R., 1948, *Industr. and Engng Chem.*, **40**, 846.

Barrell, H., 1950, *J. Sci. Instrum.*, **27**, 151.

Baxter, I. G., 1953, *J. Sci. Instrum.*, **30**, 358, 456.

BAYARD, R. T., and ALPERT, D., 1950, *Rev. Sci. Instrum.*, **21**, 672.
BECK, A. H., and BRISBANE, A. D., 1952, *Vacuum*, **2**, 137.
BELL, R. L., 1956, *J. Sci. Instrum.*, **33**, 269.
BENNETT, W. H., 1950, *J. Appl. Phys.*, **21**, 143.
BENSON, J. M., 1956, *Trans. National Symposium on Vacuum Technology*, 87.
BERRY, C. E., 1954, *J. Appl. Phys.*, **25**, 28.
BLEARS, J., 1947, *Proc. Roy. Soc.* A, **188**, 62.
BLEARS, J., and LECK, J. H., 1951 a, *J. Sci. Instrum.*, Supplement No. 1, 20; 1951 b, *Brit. J. Appl. Phys.*, **2**, 227.
BLOOMER, R. N., and HAINE, M. E., 1953, *Vacuum*, **3**, 128.
BOYD, R. L. F., and MORRIS, D., 1955, *Proc. Phys. Soc.* A, **68**, 1.
BRITISH STANDARDS INSTITUTION, 1958, *Glossary of Terms used in High Vacuum Technology*, B.S.2951 : 1958.
BROWN, G. P., 1945, *Rev. Sci. Instrum.*, **16**, 316.
CAMERON, A. E., and EGGERS, D. F., 1948, *Rev. Sci. Instrum.*, **19**, 605.
CHARPENTIER, D. E., 1956, *Trans. National Symposium on Vacuum Technology*, 114.
COOK, D. B., and DANBY, C. J., 1953, *J. Sci. Instrum.*, **30**, 238.
DAVIS, R. H., and DIVATIA, A. S., 1954, *Rev. Sci. Instrum.*, **25**, 1193.
DAYTON, B. B., 1948, *Industr. and Engng Chem.*, **40**, 778.
DAYTON, B. B., *et al.*, 1955, *Trans. National Symposium on Vacuum Technology*, 91.
DIBELER, V. H., and CORDERO, F., 1951, *J. Res. Nat. Bur. Stands*, **46**, 1.
DOWNING, J. R., and MELLEN, G. L., 1946, *Rev. Sci. Instrum.*, **17**, 218.
DUSHMAN, S., 1949, *Vacuum Technique* (New York : Wiley).
FOSTER, J. S., 1953, *Report* UCRL 2312, Radiation Lab., University of California.
FOSTER, J. S., *et al.*, 1953, *Rev. Sci. Instrum.*, **24**, 388.
GAEDE, W., 1947, *Z. Naturf.*, **2A**, 233.
GALE, A. J., 1956, *Trans. National Symposium on Vacuum Technology*, 12.
GOULD, C. L., 1956, *Trans. National Symposium on Vacuum Technology*, 39.
GREENOUGH, M. L., and WILLIAMS, W. E., 1951, *J. Res. Nat. Bur. Stands*, **46**, 5.
HAGSTRUM, H. D., 1953, *Rev. Sci. Instrum.*, **24**, 1122.
HALL, L. D., 1958, *Rev. Sci. Instrum.*, **29**, 367.
HERB, R. G., *et al.*, 1953, *Phys. Rev.*, **93**, 926.
HICKMAN, K. C. D., 1940, *J. Appl. Phys.*, **11**, 303; 1951, *Rev. Sci. Instrum.*, **22**, 141.
HICKMAN, K. C. D., and KINSELLA, J. J., 1956, *Trans. National Symposium on Vacuum Technology*, 52.
HIPPLE, J. A., *et al.*, 1949, *Phys. Rev.*, **76**, 1877.
HOBSON, J. P., and REDHEAD, P. A., 1957, *Report on 17th Annual Conference on Physical Electronics, Massachusetts Inst. of Technology*; 1958, *Canad. J. Phys.*, **36**, 271.
JACOBS, R. B., and ZUHR, H. F., 1947, *J. Appl. Phys.*, **18**, 34.
JAECKEL, R., 1947, *Z. Naturf.*, **2a**, 666.
KATZENSTEIN, H. S., and FRIEDLAND, S. S., 1955, *Rev. Sci. Instrum.*, **26**, 324.
KENNEDY, P. B., and SMITH, H. R., 1955, *Trans. National Symposium on Vacuum Technology*, 80.
KLUMB, H., *et al.*, 1953, *Naturwissenschaften*, **40**, 196.
LANDER, J. J., 1950, *Rev. Sci. Instrum.*, **21**, 672.
LATHAM, D., *et al.*, 1951, *Vacuum*, **1**, 97; 1952, *Ibid.*, **2**, 33.
LECK, J. H., 1957, *Pressure Measurement in Vacuum Systems* (London : Institute of Physics).
LIFSHITZ, I. M., and ROSENSWEIG, L. N., 1952, *J. Tech. Phys., Moscow*, **22**, 1362.
LILLY, J. C., *et al.*, 1947, *J. Appl. Phys.*, **18**, 613.
LITTING, C. N. W., and TAYLOR, W. K., 1952, *Proc. Instn Elect. Engrs*, Pt. IV, **99**, 241 (I.E.E. Monograph No. 36).
MALTER, L., and MARCUVITZ, N., 1938, *Rev. Sci. Instrum.*, **9**, 92.
MARTIN, A. J., 1955, *J. Sci. Instrum.*, **32**, 400.
MELLEN, G. L., 1948, *Industr. and Engng Chem.*, **40**, 787.
METSON, G. H., 1951, *Brit. J. Appl. Phys.*, **2**, 46.
MILLAR, W., 1956, British Patent Application No. 12242/56.

MILNER, C. J., 1951, *J. Sci. Instrum.*, Supplement No. 1, 29.
MOODY, R. E., 1956, *Trans. National Symposium on Vacuum Technology*, 114.
MORGAN, W. A., *et al.*, 1953, *American Chemical Society Symposium on Process Instrumentation, Chicago*, Sept. 1953.
NÖLLER, H. G., 1955, *Vacuum*, **5**, 59; 1956, *Trans. National Symposium on Vacuum Technology*, 57.
NÖLLER, H. G., and KUTSCHER, H., 1955, *Z. angew. Phys.*, **7**, 218.
OATLEY, C. W., 1954, *Brit. J. Appl. Phys.*, **5**, 358.
OPSTELTEN, J. J., and WARMOLTZ, N., 1955, *Appl. Sci. Res., Hague*, B, **4**, 329.
OPSTELTEN, J. J., *et al.*, 1956, *Appl. Sci. Res., Hague*, B, **6**, 129.
POLLARD, J., 1951, *S.E.R.L. Technical Journal*, **1**, 66; 1953, *Rev. Sci. Instrum.*, **24**, 996.
POLLARD, J., *et al.*, 1948, *J. Sci. Instrum.*, **25**, 401.
POWER, B. D., and CRAWLEY, D. J., 1954, *Vacuum*, **4**, 415.
POWER, B. D., and KENNA, R. A., 1955, *Vacuum*, **5**, 35.
PRESSEY, D. C., 1953, *J. Sci. Instrum.*, **30**, 20.
REDHEAD, P. A., 1952, *Canad. J. Phys.*, **30**, 1; 1957, *Report on 17th Annual Conference on Physical Electronics, Massachusetts Inst. of Technology*; 1958, *Report on 18th Annual Conference on Physical Electronics, Massachusetts Inst. of Technology*.
REDHEAD, P. A., and CROWELL, C. R., 1953, *J. Appl. Phys.*, **24**, 331.
REIMANN, A. L., 1952, *Vacuum Technique* (London: Chapman and Hall), p. 83.
SCHWARZ, H. J., 1956, *Trans. National Symposium on Vacuum Technology*, 37.
SMITH, H. R., 1955, *Trans. National Symposium on Vacuum Technology*, 22.
SOMMER, H., *et al.*, 1951, *Phys. Rev.*, **82**, 697.
STARK, D., 1957, *Brit. J. Appl. Phys.*, **8**, 210.
STEINHERZ, H. A., 1956, *Trans. National Symposium on Vacuum Technology*, 62.
SWARTZ, J. C., 1955, *Trans. National Symposium on Vacuum Technology*, 83.
THEES, R., 1955, *Vacuum*, **5**, 25.
THOMAS, H. A., *et al.*, 1946, *Rev. Sci. Instrum.*, **17**, 368.
VACCA, R. H., 1956, *Trans. National Symposium on Vacuum Technology*, 93.
VAN ATTA, C. M., 1956, *Trans. National Symposium on Vacuum Technology*, 62.
VARADI, P. F., and SEBASTYEN, L. G., 1956, *J. Sci. Instrum.*, **33**, 392.
VARNERIN, L. J., and CARMICHAEL, J. H., 1955 a, *J. Appl. Phys.*, **26**, 782; 1955 b, *Bull. Amer. Phys. Soc.*, **30**, 36; 1957, *J. Appl. Phys.*, **28**, 913.
VENEMA, A., 1954, *Vacuum*, **4**, 272.
VOGT, A., 1954, British Patent No. 709119.
VON UBISCH, H., 1952, *Analyt. Chem.*, **24**, 931.
WHITE, W. C., and HICKEY, J. S., 1948, *Electronics*, **21**, 100.
WILEY, W. C., and MCLAREN, I. H., 1955, *Rev. Sci. Instrum.*, **26**, 324.
WOLFF, M. M., and STEPHENS, W. E., 1953, *Rev. Sci. Instrum.*, **24**, 616.
WORCESTER, W. G., and DOUGHTY, E. G., 1946, *Trans. Amer. Inst. Elect. Engrs*, **65**, 948.
YOUNG, J. R., 1956, *J. Appl. Phys.*, **27**, 926.

THE PROPAGATION OF ELASTIC WAVES IN CRYSTALS AND OTHER ANISOTROPIC MEDIA

By M. J. P. MUSGRAVE
National Physical Laboratory, Teddington, Middlesex

CONTENTS

	PAGE
§ 1. Introduction	75
§ 2. General exposition	75
2.1. Velocity, inverse (slowness) and wave surfaces	75
2.2. Generalized Hooke's law	76
2.3. Equations of motion	78
2.4. Plane waves	78
2.5. General characteristics of propagation	80
§ 3. Media of various symmetries	81
3.1. Hexagonal symmetry	81
3.2. Calculated characteristics of propagation	82
3.3. Experimentally observed characteristics of propagation	86
§ 4. Reflection and refraction	88
4.1. Boundary conditions at a plane interface	88
4.2. Critical angles and surface waves	88
4.3. Determination of possible reflected and refracted waves	89
§ 5. Surface waves of Rayleigh and Love type	90
5.1. General discussion	90
5.2. Hexagonal and cubic media	91
5.3. Monoclinic and orthorhombic media	91
§ 6. Stationary waves	92
§ 7. Propagation of elastic waves in polycrystalline aggregates	92
7.1. Relation of single crystal to the aggregate	92
7.2. Elastic anisotropy due to preferred orientation	93
7.3. Elastic anisotropy due to laminated structure	93
7.4. Problems of attenuation	94
§ 8. Relation between theory and experiment	94
Acknowledgments	95
References	95

Abstract. The general principles governing the behaviour of plane elastic waves in aeolotropic media are developed and the significance of the velocity, slowness and wave surfaces is indicated. Sections of these surfaces for media of hexagonal and cubic symmetry are illustrated and discussed from a theoretical and an experimental standpoint.

The general problem of reflection and refraction which arises when a plane wave is incident upon a plane interface between two arbitrarily oriented aeolotropic media is discussed and exemplified by a simple case involving two copper crystals.

The conditions under which surface waves of Rayleigh and Love types may be propagated are discussed.

Problems of elastic wave propagation in anisotropic aggregates are mentioned and briefly examined.

The article concludes with a consideration of the main assumptions relating theory and experiment.

§ 1. Introduction

The study of the propagation of waves in elastic media has a history which is closely associated with the development of the wave theory of light (Whittaker 1951) and therefore commences with the publication of Huygens' *Treatise* (1690). The objections which Newton (1704) raised—quite validly, since only longitudinal waves were conceived at that time—prevailed for over a hundred years until Young in 1817 suggested that light waves might possess transverse displacement vectors. This idea was brilliantly developed by Fresnel (1827) into the theory of crystal optics in which the aether was considered as an elastic solid.

Throughout the nineteenth century, various types of elastic solid were proposed to represent the luminiferous aether. Hamilton (1832), Cauchy (1836), Green (1842), McCullagh (1848) and others developed the analyses appropriate to these models; however, their interest was principally focused on the phenomena of light which were adequately explained by postulating a solid which would sustain transverse vibrations only or, if it could sustain longitudinal vibrations, in which these were imperceptible. Waves possessing a longitudinal or dilatational component of displacement seem to have been more a source of embarrassment than of interest though Blanchet (1840) made some investigations of a general aeolotropic elastic medium.

At the end of the nineteenth century, Christoffel (1877) wrote a long memoir about the propagation of elastic waves in aeolotropic media and Lord Kelvin developed part of the necessary theory in his Baltimore lectures (1904) and described it as "a fine subject for investigation" before leaving it for the discussion of optical phenomena. It is clear that each of them possessed a full qualitative understanding of the subject but, in their day, means were lacking to launch or to detect mechanical vibrations of high frequency. Consequently, no observations were available on disturbances whose wavelengths were so small in comparison with sources and specimens as to exhibit the characteristics of propagation along rays or beams.

Papers by Rudski (1911) and, more recently, by Stoneley (1949) show that seismologists have constantly borne in mind the possibility that elastic anisotropy might be the cause of anomalies in seismic observations. It is this interest, together with the existence of the techniques developed in recent years for propagating elastic waves of ultrasonic frequency from sources whose dimensions are many times the wavelength, that has called for a quantitative ray theory of elastic waves in aeolotropic solids.

As might be expected, many of the characteristics of the behaviour of light waves in crystals are also exhibited by elastic waves; some such effects have been experimentally verified, others await the interest of those able to improve the techniques of propagating and detecting elastic waves.

§ 2. General Exposition

2.1. *Velocity, Inverse (Slowness) and Wave Surfaces*

If an unpolarized point source of elastic vibration is located in an isotropic medium, the disturbance spreads outwards on two concentric spherical fronts;

that which travels faster is associated with longitudinal displacement, the slower with transverse displacement. These two spheres comprise the *elastic wave surface* and Huygens' construction quickly shows that the path along which energy is carried by a plane wave of any normal coincides with that normal This is a consequence of the fact that the tangent plane at any point of a sphere is perpendicular to the radius vector. When a wave front is non-spherical, the tangent plane at any point is not, in general, perpendicular to the radius vector and the path of energy flux or ray is thus oblique to the normal of the associated plane wave element. This latter condition prevails for the *three* sheets of a general elastic wave surface in an aeolotropic medium which is the mathematical envelope at unit time of all the possible plane waves in the medium which simultaneously passed through a single point, the origin, at zero time.

In order to calculate such a surface we first examine the characteristics of plane waves in the medium. We shall find there exist three discrete phase velocities and associated displacement vectors for each possible wave normal. The locus of the extremities of the phase velocity vectors forms a *velocity surface* of three sheets. When this surface is inverted with respect to a sphere we obtain an inverse or slowness surface of three sheets which is polar reciprocal to the wave surface. This geometrical relation allows the calculation of the wave surface to be effected. If this process were performed for the case of an isotropic medium, we should find that the two slower phase velocities were equal so that the velocity surface consisted of two spheres, the smaller consisting of two coincident sheets ; the *inverse surface* would be two spheres, the larger consisting of two coincident sheets ; the wave surface would be identical with the velocity surface.

The wave surface possesses the clearest physical significance of the three surfaces since it represents the fronts on which a disturbance would be observed to spread and is also the appropriate surface for the application of Huygens' principle. However, as Synge (1957) has pointed out, mathematically the most important of the three is the inverse surface or slowness surface. The principal reasons for this are that its equation of the sixth degree is easily obtainable, its polar reciprocal relation to the wave surface allows many properties of that surface to be immediately inferred and the concept of slowness is the most convenient to use in problems of reflection and refraction.

(Note. The term ' slowness ' was first used by Hamilton and is the traditional term for a quantity having dimensions $[L^{-1}T]$; the inverse vector in this article is proportional to the slowness but has dimensions $[LT^{-1}]$. The two vectors are therefore not identical but, of course, in their geometrical properties the slowness and inverse surfaces are the same.)

In the following pages, the interesting though somewhat complicated characteristics of the propagation of elastic waves in aeolotropic media are discussed in terms of these three surfaces.

2.2. *Generalized Hooke's Law*

If we take orthogonal axes of reference Ox, Oy, Oz in a general aeolotropic medium, we may, following the classical theory of elasticity, relate stresses to strains by the set of equations

$$\widehat{xx} = c_{11}u_x + c_{12}v_y + c_{13}w_z + c_{14}(v_z + w_y) + c_{15}(w_x + u_z) + c_{16}(u_y + v_x), \quad \ldots\ldots(1)$$

$$\widehat{yy} = c_{12}u_x + c_{22}v_y + c_{23}w_z + c_{24}(v_z + w_y) + c_{25}(w_x + u_z) + c_{26}(u_y + v_x), \quad \ldots\ldots(2)$$

$$\widehat{zz} = c_{13}u_x + c_{23}v_y + c_{33}w_z + c_{34}(v_z + w_y) + c_{35}(w_x + u_z) + c_{36}(u_y + v_x), \quad \ldots\ldots(3)$$

$$\widehat{yz} = \widehat{zy} = c_{14}u_x + c_{24}v_y + c_{34}w_z + c_{44}(v_z + w_y) + c_{45}(w_x + u_z) + c_{46}(u_y + v_x), \quad \ldots\ldots(4)$$

$$\widehat{zx} = \widehat{xz} = c_{15}u_x + c_{25}v_y + c_{35}w_z + c_{45}(v_z + w_y) + c_{55}(w_x + u_z) + c_{56}(u_y + v_x), \quad \ldots\ldots(5)$$

$$\widehat{xy} = \widehat{yx} = c_{16}u_x + c_{26}v_y + c_{36}w_z + c_{46}(v_z + w_y) + c_{56}(w_x + u_z) + c_{66}(u_y + v_x), \quad \ldots\ldots(6)$$

where $c_{rs} = c_{sr}$ $(r, s = 1, 2, \ldots, 6)$ are the elastic constants of the medium referred to the axes of reference and the displacement of the point (x, y, z) is (u, v, w), letter suffixes here denoting differentiation.

Before proceeding further it is appropriate to state two assumptions which are involved in relations (1)–(6) and which have been challenged in recent papers as inadequately representing the behaviour of certain media when these undergo small deflections.

Firstly, Laval (1951), LeCorre (1953) and Raman and Viswanathan (1955) have put forward the proposition that under certain dynamic conditions it is not permissible to consider rotations as occurring bodily. They consider that rotational strain may occur within the body and that each element of volume is under the influence of shear stresses which may give rise to resultant volume couples proportional to $\widehat{yz}$–$\widehat{zy}$, etc. Clearly, this results in the stress and strain tensors becoming antisymmetric and the total possible number of independent constants is increased to forty-five. Joel and Wooster (1957) have pointed out that there is good reason to expect the classical theory of elasticity to prove inadequate for those classes of symmetry which contain optically active crystals. Irrefutable experimental evidence of this possible inadequacy has yet to appear and we shall not at present consider the consequences of it further.†

Secondly, Makinson (1955) has suggested that for certain complex molecular crystals or oriented polymers the relation $c_{rs} = c_{sr}$ may be invalid. This strikes at the very basis of the classical theory since if $c_{rs} \neq c_{sr}$ the strain energy function does not define a unique state of deformation. Some of the consequences of thus assuming the existence of thirty-six possible independent constants are interesting but experimental evidence is again wanting.

Accordingly we shall continue this discussion in the terms of the classical theory of elasticity bearing in mind that, for all the frequencies at which it is at present possible to launch and detect elastic waves, propagation takes place under

† *Note added in proof*: Discussion of the number of elastic constants appropriate to crystals continues, see recent letters to *Nature, Lond.*, **182**, by R. S. Krishnan, *et al.*, p. 518, F. A. McClintock, *et al.*, p. 652, and N. Joel and W. A. Wooster, pp. 1078, 1149.

conditions of constant entropy and the adiabatic values of constants are appropriate. No treatment of the thermo-elastic equations of motion for an aeolotropic solid seems yet to have appeared, but useful discussions of these equations for isotropic solids have been given by Deresiewicz (1957) and Chadwick and Sneddon (1958).

The relations (1)–(6) involve twenty-one constants which in a medium of triclinic symmetry will be independent. However, in any medium of higher symmetry they will be functions of a smaller set of independent constants and the orientation of the reference axes with respect to the symmetry axes of the medium; by suitable choice of the reference axes relations (1)–(6) may be much simplified. It will be assumed that the axes of reference have been so chosen.

2.3. *Equations of Motion*

The equations of motion for an elastic medium from which body forces are absent are

$$\frac{\partial(\widehat{xx})}{\partial x}+\frac{\partial(\widehat{xy})}{\partial y}+\frac{\partial(\widehat{xz})}{\partial z}=\rho\ddot{u}, \qquad \ldots\ldots(7)$$

and two similar, (8), (9), where ρ is the density of the medium. Substitution of the forms (1)–(6) yields a set of three simultaneous partial differential equations for the components of displacement (u, v, w) as functions of the space variables, time and constants of the medium.

2.4. *Plane Waves*

Following Kelvin (1904), let us assume a plane wave solution to the equations of motion of the form

$$(u, v, w)=(p, q, r)\exp\frac{i2\pi}{\lambda}[lx+my+nz-v(l, m, n)t], \qquad \ldots\ldots(10)$$

where (p, q, r) is a displacement vector of unit amplitude associated with a plane wave of normal (l, m, n), phase velocity $v(l, m, n)$ and wavelength λ. On substitution, we obtain

$$\begin{bmatrix} A-\rho v^2 & \alpha\beta & \gamma\alpha \\ \alpha\beta & B-\rho v^2 & \beta\gamma \\ \gamma\alpha & \beta\gamma & C-\rho v^2 \end{bmatrix}\begin{bmatrix} p \\ q \\ r \end{bmatrix}=0, \qquad \ldots\ldots(11)$$

where v, the phase velocity, is written without direction cosines since the displacement component v has been eliminated from the notation. A, B, C, $\beta\gamma$, $\gamma\alpha$, $\alpha\beta$ are each functions of (l, m, n) and c_{rs} and represent stiffnesses; in notation due to Christoffel, they appear as Γ_{ii} and Γ_{ij} $(i, j=1, 2, 3)$. They are given by

$$A=\Gamma_{11}=l^2c_{11}+m^2c_{66}+n^2c_{55}+2mnc_{56}+2nlc_{15}+2lmc_{16}, \qquad \ldots\ldots(12)$$

and two similar, (13), (14),

$$\beta\gamma=\Gamma_{23}=l^2c_{56}+m^2c_{24}+n^2c_{34}+mn(c_{22}+c_{44})+nl(c_{36}+c_{45})+lm(c_{25}+c_{46}), \qquad \ldots\ldots(15)$$

and two similar, (16), (17).

Non-zero solutions of (p, q, r) occur for values of v which satisfy the determinantal cubic equation

$$\begin{vmatrix} A-\rho v^2 & \alpha\beta & \gamma\alpha \\ \alpha\beta & B-\rho v^2 & \beta\gamma \\ \gamma\alpha & \beta\gamma & C-\rho v^2 \end{vmatrix} = 0. \qquad \ldots\ldots(18)$$

The symmetry of this determinant implies the existence of three positive roots ρv_i^2 whose associated displacement vectors are mutually orthogonal and may be written

$$p_i : q_i : r_i = \frac{\alpha}{\rho v_i^2 - A + \alpha^2} : \frac{\beta}{\rho v_i^2 - B + \beta^2} : \frac{\gamma}{\rho v_i^2 - C + \gamma^2}. \qquad \ldots\ldots(19)$$

Examination of the displacement vectors represented by (19) shows that, in general, none is truly longitudinal or transverse; however, in most crystals, the vector associated with the largest velocity is nearly longitudinal and the other two are consequently nearly transverse. It is therefore convenient to designate the waves quasi-longitudinal ($i = L$) or quasi-transverse ($i = T_1$ or T_2), as shown in figure 1.

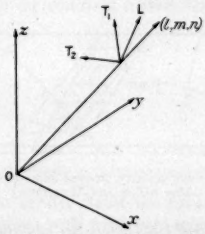


Figure 1. Orthogonal triad of displacement vectors associated with any wave normal (l, m, n).

Clearly, the deviation δ_i of any of these displacement vectors from the wave normal is given by

$$\cos\delta_i = lp_i + mq_i + nr_i. \qquad \ldots\ldots(20)$$

The plot of $v_i(l, m, n)$ yields the velocity surface which possesses three sheets.

The geometrical relation between a velocity, an inverse and a wave point is shown in figure 2(*a*) and (*b*). Quantitative expressions for the several line segments have been given by Musgrave (1954 a) and these yield coordinates for a wave point

$$\xi_i = \frac{lk}{\rho v_i} + \frac{L_i}{2\rho v_i} = \frac{1}{\rho}(lk + \tfrac{1}{2}L_i)s_i, \qquad \ldots\ldots(21)$$

$$\eta_i = \frac{mk}{\rho v_i} + \frac{M_i}{2\rho v_i} = \frac{1}{\rho}(mk + \tfrac{1}{2}M_i)s_i, \qquad \ldots\ldots(22)$$

$$\zeta_i = \frac{nk}{\rho v_i} + \frac{N_i}{2\rho v_i} = \frac{1}{\rho}(nk + \tfrac{1}{2}N_i)s_i, \qquad \ldots\ldots(23)$$

where k is some elastic constant and determines the sphere of inversion between the velocity and inverse surfaces, s_i is the modulus of the slowness vector $\mathbf{s}_i$, i.e.

$s_i = |\mathbf{s}_i| = 1/v_i$, and (L_i, M_i, N_i) are the components of IP where

$$L_i = \frac{p_i^2}{\alpha^2}\left[\frac{\partial \alpha^2}{\partial l}(\rho v_i^2 - A) + \alpha^2 \frac{\partial A}{\partial l}\right] + \frac{q_i^2}{\beta^2}\left[\frac{\partial \beta^2}{\partial l}(\rho v_i^2 - B) + \beta^2 \frac{\partial B}{\partial l}\right] + \frac{r_i^2}{\gamma^2}\left[\frac{\partial \gamma^2}{\partial l}(\rho v_i^2 - C) + \gamma^2 \frac{\partial C}{\partial l}\right], \quad \ldots\ldots(24)$$

and M_i and N_i are similar functions (25) and (26) but with m or n in place of l.

It is convenient to write direction cosines

$$\lambda_i : \mu_i : \nu_i = \frac{L_i : M_i : N_i}{[L_i^2 + M_i^2 + N_i^2]^{1/2}}, \quad \ldots\ldots(27)$$

when ϵ, shown in figure 2, is given by

$$\cos \epsilon_i = l\lambda_i + m\mu_i + n\nu_i. \quad \ldots\ldots(28)$$

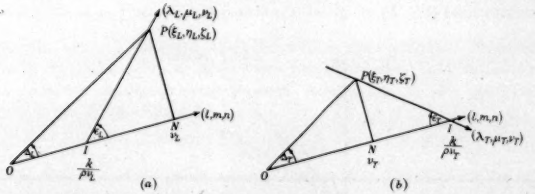


Figure 2. Geometrical relation between velocity, inverse and wave points ; (*a*) $v > k/\rho v$, (*b*) $v < k/\rho v$. (*After* Miller and Musgrave.)

2.5. *General Characteristics of Propagation*

From the results of the analysis which has been sketched, it is possible to make some general assertions about the behaviour of elastic waves when propagated in crystals.

When a plane area of any aeolotropic medium is mechanically excited in phase at a frequency whose corresponding wavelength in the medium is small, energy is propagated from the source on a diverging wave front of small but finite curvature ; the consequent highly directional transmission may be understood in terms of the ray theory of sound.

Thus, the given plane area, e.g. the plane face of a crystal specimen, defines a single wave normal and a triad of possible displacement vectors, each of which may be propagated with a discrete phase velocity. When a periodic excitation of any displacement is applied, this displacement will be resolved into its components along the possible displacement vectors and each component vibration will be transmitted at its own phase velocity. In general, the three possible waves will be quasi-longitudinal and quasi-transverse and none will have a ray coincident with the wave normal. Consequently, if the orientation of the wave normal is arbitrary, we should expect triple extraordinary refraction of a truly longitudinal or transverse wave incident upon the face of a crystal, as in figure 3.

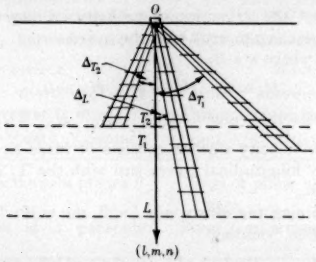


Figure 3. Beams within which vibrational energy, radiated from a finite source, is confined in passage through an aeolotropic medium. The broken lines parallel to the excited surface represent the planes which each type of vibration has just reached after time t; the distances OL, OT_1 and OT_2 are directly proportional to the three phase velocities for direction (l, m, n). Note: in general these beams are not coplanar.

It follows that energy may travel back and forth across a specimen with parallel sides along a path oblique to the common surface normal, as shown in figure 4.

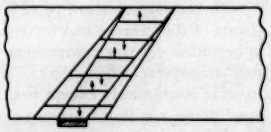


Figure 4. Reflection back and forth oblique to the wave normal defined by perpendicular incidence.

We will now examine in greater detail the propagation of elastic waves in a medium of hexagonal symmetry with particular reference to crystalline zinc.

§ 3. Media of Various Symmetries

3.1. *Hexagonal Symmetry*

When referred to axes Ox, Oy lying in the basal plane and Oz normal to it, the elastic properties of hexagonal media may be described in terms of five independent constants c_{11}, c_{12}, c_{13}, c_{33}, c_{44}.

Writing, for convenience,

$$c_{11}-c_{44} = a, \qquad \ldots\ldots(29)$$

$$c_{11}-c_{12}-2c_{44} = c, \qquad \ldots\ldots(30)$$

$$c_{13}+c_{44} = d, \qquad \ldots\ldots(31)$$

$$c_{33}+c_{44} = h, \qquad \ldots\ldots(32)$$

$$\rho v^2-c_{44} = H, \qquad \ldots\ldots(33)$$

the determinantal cubic for the velocity surface may be expanded in the form

$$H^3-\{n^2h+(1-n^2)(a+\tfrac{1}{2}c)\}H^2 + (1-n^2)[(1-n^2)\tfrac{1}{2}ac+n^2\{h(a+\tfrac{1}{2}c)-d^2\}]H-\tfrac{1}{2}n^2(1-n^2)^2c(ah-d^2) = 0, \qquad \ldots\ldots(34)$$

when it is evident that the velocities of propagation vary with n only and the surface possesses circular symmetry about the z-axis.

In the basal plane where $n = 0$,

$$H_{\mathrm{L}} = a, \quad H_{\mathrm{T}_1} = \tfrac{1}{2}c, \quad H_{\mathrm{T}_2} = 0, \qquad \ldots\ldots(35)$$

or

$$\rho v^2 = c_{11}, \quad \rho v_{\mathrm{T}_1}{}^2 = \tfrac{1}{2}(c_{11} - c_{12}), \quad \rho v_{\mathrm{T}_2}{}^2 = c_{44}, \qquad \ldots\ldots(36)$$

and the section of the velocity surface consists of three circles; associated with the L section are truly longitudinal waves and with the T_1 and T_2 sections, truly transverse waves.

In any plane containing the z-axis, say $l = 0$, the cubic factorizes into a linear and a quadratic equation, yielding

$$H_{\mathrm{T}_1} = \tfrac{1}{2}m^2 c, \quad H_{\mathrm{L,T}_2} = \tfrac{1}{2}[m^2 a + n^2 h \pm \{(m^2 a + n^2 h)^2 - 4m^2 n^2 (ah - d^2)\}^{1/2}], \qquad \ldots\ldots(37)$$

with associated displacement vectors

$$p_{\mathrm{T}_1} = 1, \; q_{\mathrm{T}_1} = 0 = r_{\mathrm{T}_1}; \quad p_{\mathrm{L,T}_2} = 0, \frac{q_{\mathrm{L,T}_2}}{r_{\mathrm{L,T}_2}} = \frac{mnd}{H_{\mathrm{L,T}_2} - m^2 a} = \frac{H_{\mathrm{L,T}_2} - n^2 h}{mnd}. \qquad \ldots\ldots(38)$$

Equations (37) correspond with the factorization of (34) given by Synge (1957); he also shows that the T_1 sheets of the inverse and wave surfaces which correspond with $H_{\mathrm{T}_1} = \frac{1}{2}(1 - n^2)c$ are spheroidal. From expressions (38) it may be seen that the T_1 wave is always truly transverse while the L and T_2 waves are quasi-longitudinal and quasi-transverse respectively except for the special cases of waves whose normals lie in the basal plane $z = 0$ or normal to it.

The components (L_i, M_i, N_i) of the line segments between inverse and wave points which correspond with the velocity points defined by (37) are

$$L_{\mathrm{T}_1} = 0 = N_{\mathrm{T}_1}, \quad M_{\mathrm{T}_1} = mc; \qquad \ldots\ldots(39)$$

$$L_{\mathrm{L,T}_2} = 0, \quad M_{\mathrm{L,T}_2} = \frac{2q_{\mathrm{L,T}_2}{}^2}{m} H_{\mathrm{L,T}_2}, \quad N_{\mathrm{L,T}_2} = \frac{2r_{\mathrm{L,T}_2}{}^2}{n} H_{\mathrm{L,T}_2}; \qquad \ldots\ldots(40)$$

whence

$$\cos \epsilon_{\mathrm{T}_1} = m, \quad \cos \epsilon_{\mathrm{L,T}_2} = \left[\frac{q_{\mathrm{L,T}_2}{}^4}{m^2} + \frac{r_{\mathrm{L,T}_2}{}^4}{n^2}\right]^{-1/2}.$$

3.2. *Calculated Characteristics of Propagation*

Using the above expressions and some published values for the elastic constants of zinc, Musgrave (1954 b) has computed the velocity, inverse and wave surfaces for a zinc crystal and also given graphs showing the variation of δ_{L}, the deviation of the quasi-longitudinal wave, and of Δ_i (i = L, T_1 or T_2), the deviation of the ray from the wave normal. These curves are reproduced in figures 5(*a*), (*b*), (*c*), (6) and (7).

At this stage it is appropriate to point out some geometrical relations which exist between the inverse and wave surfaces and which are exemplified by the sections of figures 5(*b*) and (*c*).

As already stated, the inverse and wave surfaces are polar reciprocal to each other and the principle of duality (of projective geometry) ensures the following projective correspondences which relate the tangent planes and points of the one surface to the points and tangent planes of the other.

Projective Correspondences which follow from the Polar Reciprocal Relation between Inverse and Wave Surfaces

Inverse surface	Wave surface
Tangent plane P : normal $(\xi:\eta:\zeta)$	Point p: coordinates (ξ, η, ζ)
Point Q : radius vector $(l:m:n)$	Tangent plane q : normal $(l:m:n)$
Tangent plane $P_s^{(N)}$ with N contacts at points Q_s	Point $p_s^{(N)}$ with N tangent planes q_s
Point $Q_s^{(N)}$ with N tangent planes P_s	Tangent plane $q_s^{(N)}$ with N contacts at p_s
Tangent plane $\mathfrak{P}$ at a parabolic point $\mathfrak{Q}$ (At a parabolic point on a surface, one or both of the principal curvatures is zero)	Point $\mathfrak{p}$ on a cuspidal edge

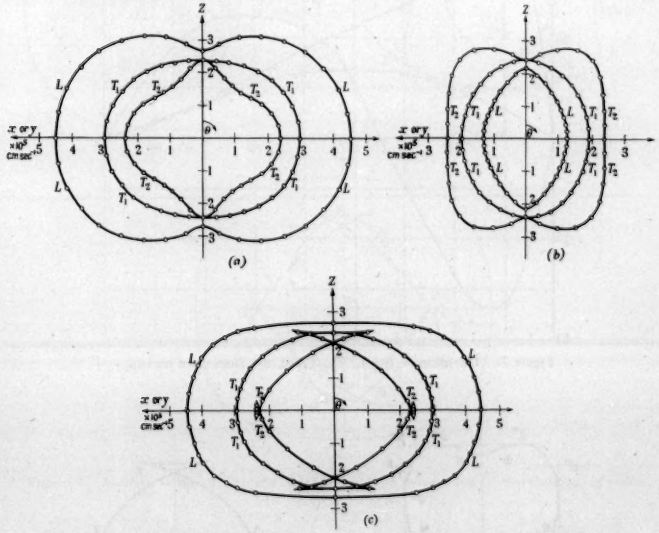


Figure 5. Sections of (*a*) velocity surface, (*b*) inverse surface and (*c*) wave surface, for zinc.

Some of these correspondences are exemplified by the inverse and wave T_2 sheets of zinc as is shown in figures 8(*a*) and (*b*). Thus any plane of type P^{II} corresponds with a point of type p^{II} and the plane P^{∞} making contact at all points Q^{∞} of a circle corresponds with the point p^{∞} which has a circular cone of tangency ; also tangent planes $\mathfrak{P}$ having contact at $\mathfrak{Q}$ correspond with points $\mathfrak{p}$.

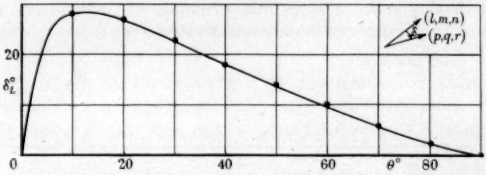


Figure 6. Deviation δ_L of quasi-longitudinal vector from wave normal.

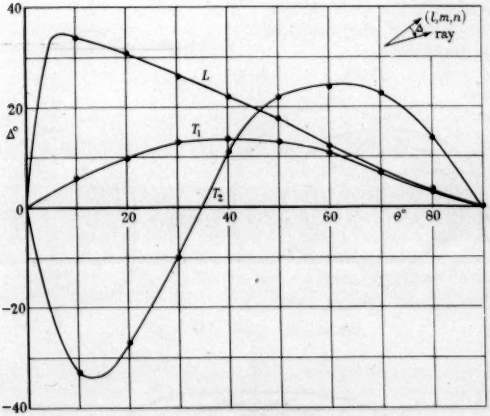


Figure 7. Deviation Δ_i (i = L, T_1, T_2) of rays from wave normal.

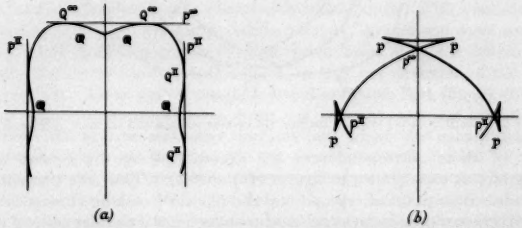


Figure 8. (*a*) Inverse section and (*b*) wave section (zinc) illustrating projective correspondences.

The sections of velocity, inverse and wave surfaces for a (1, 1, 0) plane of the cubic crystal nickel are also shown in figures 9(*a*), (*b*), (*c*), which are taken from an account of waves in cubic crystals by Miller and Musgrave (1956). The intersection of the inverse T_A and T_B sheets at (l, m, n) equal to (1, 1, 1) is a conical point Q^{∞} possessing a circular cone of tangency and corresponds with a plane q^{∞} having contact at all points of a circle on the wave surface.

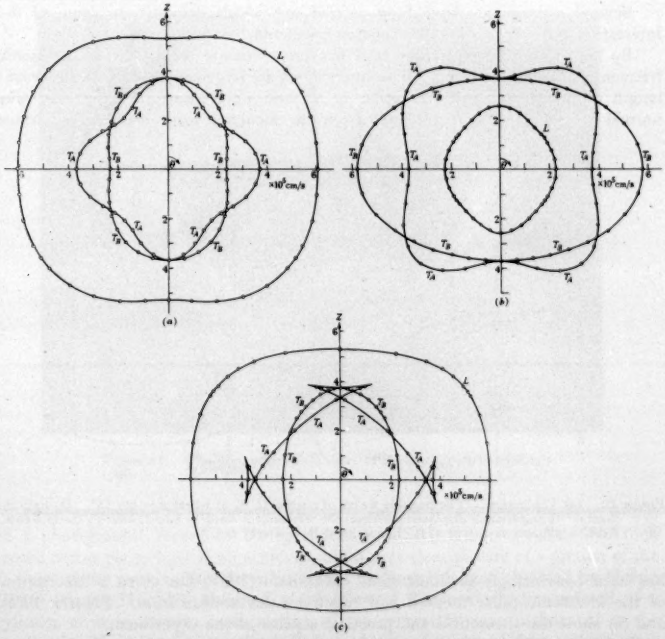


Figure 9. Sections in a (1, 1, 0) plane of (*a*) velocity, (*b*) inverse and (*c*) wave surface of nickel. (*After* Miller and Musgrave.)

The physical interest of these correspondences lies in the fact that the existence of P^{∞}, p^{∞} and Q^{∞}, q^{∞} indicates that the phenomena of external and internal conical refraction shoud be observed in zinc and nickel respectively.

From figure 7 it is apparent that the deviation of the ray from the wave normal may be as much as 35° in zinc and this order of deviation also occurs in such metals as copper, nickel and iron. Such angles of deviation can cause embarrassment in practice if they are not expected or understood (Markham 1957).

Various papers have been published on problems relating to the special orientation of the wave normal (Alexandrov 1956, Borgnis 1955), the computation of elastic constants from measured velocities (Neighbours *et al.* 1952), conical refraction (Musgrave 1957 a), and cuspidal edges on wave surfaces (Musgrave 1957 b). Further references may be found in a review by Hearmon (1956).

3.3. *Experimentally Observed Characteristics of Propagation*

Several experiments have been carried out which demonstrate some of the interesting features of wave propagation predicted by theory.

By launching and receiving small packets of elastic waves having ultrasonic frequencies from transducers whose dimensions are large compared with the wavelength of propagation, the deviation of a quasi-longitudinal ray from its wave normal in a zinc crystal as predicted from the calculated form of the wave surface

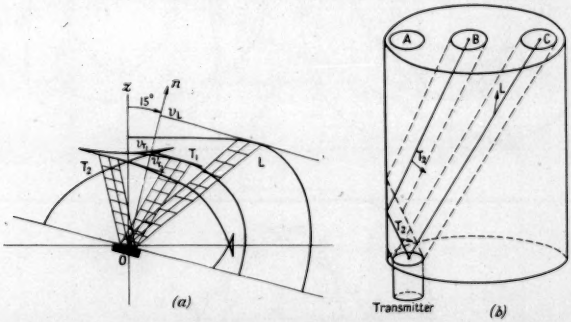


Figure 10. (*a*) Theoretical. The possible waves of normal (l, m, n) where $n = \cos 15°$. (*b*) Experimental. Longitudinal displacement from the transmitter resolves into L and T_1 components. Receiver detects no energy at A, T_2 energy at B, L energy at C.

has been convincingly demonstrated. Markham (1957) has given a description of the ultrasonic pulse method and mentions this confirmation. Figures 10(*a*) and (*b*) show the theoretical and practical aspects of the experiment.

In figure 10(*a*) are shown the three possible plane waves which may be launched from the plane face of a zinc crystal whose normal is at 15° to the zonal axis. To excite the T_1 wave, a truly transverse displacement is necessary; the L and T_2 waves, being quasi-longitudinal and quasi-transverse respectively, are both generated by the application of a truly longitudinal displacement and either wave may be detected by means of the component of its displacement vector normal to the specimen surface, although the sensitivity to the T_2 wave is, of course, smaller than to the L wave. The necessary acoustic contact may thus be preserved by an oil film and the receiver can then be readily slid from one position to another on the side of the specimen remote from the transmitter. By this means the

direction of propagation of the L wave in the zinc crystal specimen was convincingly demonstrated as shown in figure 10(*b*). A wave of much smaller amplitude and slower velocity was detected in another position of the receiver and this was almost certainly the T_2 wave which was excited and whose energy travelled along the beam reflected from the side of the specimen, the phase velocity along the axis of the specimen remaining constant (see §4, equation (41)).

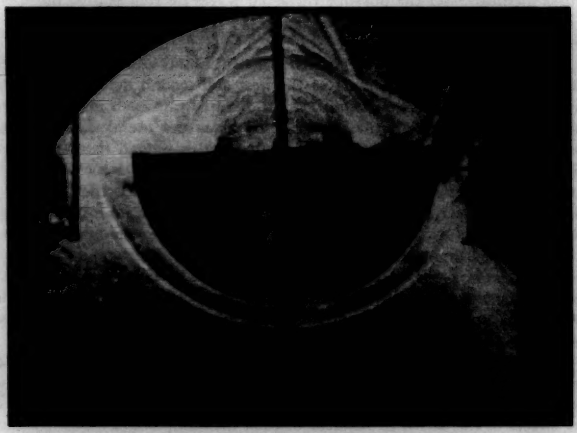


Figure 11. The non-spherical L sheet of Pertinax. (*After* Helbig.)

Ideal verification of the form of the wave surface in any given medium would be a photographic record of the spread of a disturbance in a solid. No such record seems yet to have been achieved, but a very clear picture of a section of the L sheet of the wave surface for a hexagonal medium has been obtained by Helbig (1956) (figure 11). He obtained a specimen of Pertinax which consisted of a cylinder of semi-circular cross section whose generators were orthogonal to the zonal axis of hexagonal symmetry. By launching a disturbance from a line source along the axis of this cylinder when immersed in a fluid, he was able to photograph the L sheet in the fluid after it had left the cylinder. The departure from sphericity is very clearly demonstrated.

Experimental verification of the prediction of internal conical refraction in cubic crystals has been performed (de Klerk and Musgrave 1955).

Rotation of the transverse displacement vector of a wave propagated along the axis of a twisted crystal has been observed by Alexandrov and Khanmov-Mal'kov (1956). A quarter-wave plate for elastic waves has been cut from quartz and circularly polarized waves have thus been produced (Alexandrov 1957).

§ 4. Reflection and Refraction

4.1. *Boundary Conditions at a Plane Interface*

Suppose a plane elastic wave is incident upon a plane boundary separating two aeolotropic media and let the incident and possible reflected waves be represented by

$$\mathbf{p}_I \exp i\omega(\mathbf{r}.\mathbf{s}_I - t), \quad a_i \mathbf{p}_i \exp i\omega(\mathbf{r}.\mathbf{s}_i - t) \quad (i = \mathrm{L}, \mathrm{T}_1, \mathrm{T}_2)$$

in the first medium ; let the possible refracted waves be represented by

$$a_i' \mathbf{p}_i' \exp i\omega(\mathbf{r}'.\mathbf{s}_i' - t) \quad (i = \mathrm{L}, \mathrm{T}_1, \mathrm{T}_2)$$

in the second medium. The components of unprimed and primed vectors are referred to the crystallographic axes of the first and second medium respectively. Let unit vectors normal to the boundary and directed into each medium be $\mathbf{n}$ and $\mathbf{n}'$; the plane of incidence is then determined by $\mathbf{s}_I$ and $\mathbf{n}$ and makes a trace on the boundary whose unit vector referred to the first or second medium is $\mathbf{b}$ or $\mathbf{b}'$ accordingly.

At the boundary the three components of displacement and the three of stress must be continuous for the system of waves. For this condition to be upheld at all times, the phase of each wave must change at the same rate along the boundary, that is

$$\mathbf{s}_I.\mathbf{b} = \mathbf{s}_i.\mathbf{b} = \mathbf{s}_i'.\mathbf{b}'. \qquad \ldots\ldots(41)$$

Thus the trace of the slowness or inverse vector of each wave on the boundary is the same. The precise values of the slowness vectors are determined by the intersections of the lines

$$\mathbf{s} = (\mathbf{s}_I.\mathbf{b})\,\mathbf{b} + h\mathbf{n}, \qquad \ldots\ldots(42)$$

$$\mathbf{s}' = (\mathbf{s}_I.\mathbf{b})\,\mathbf{b}' + h'\mathbf{n}', \qquad \ldots\ldots(43)$$

with the sheets of the slowness surfaces of each medium, S and S′.

Geometrical considerations show that under all circumstances these intersections, real or imaginary, determine three slowness vectors associated with appropriate rays for each medium and only the amplitude of each wave remains unknown. These six amplitudes are determined by the six equations expressing the continuity of displacement and stress. When, for a given incident slowness vector, there may be found three real points of intersection of (42) with S and of (43) with S′, such that the normals to S and S′ at these points have positive components along $\mathbf{n}$ and $\mathbf{n}'$ respectively, there will be three body waves transmitted from the boundary into each medium. In this case, the six equations determining the amplitudes are linear simultaneous equations for a_i, and a_i' whose coefficients are rather involved, real functions of the elastic constants and the orientation with respect to the boundary of each medium.

4.2. *Critical Angles and Surface Waves*

For any given orientation of the boundary with respect to the first medium, there will exist critical angles of incidence for each type of wave when, owing to the magnitude of $\mathbf{s}_I.\mathbf{b}$, some or all of the points of intersection of the line (42) with S become imaginary ; similar critical conditions can prevail over the intersections of (43) with S′. The slowness vectors so determined by these imaginary

intersections have complex components and take the form

$$\mathbf{s} = (\mathbf{s}_1 . \mathbf{b})\,\mathbf{b} + (h + i\delta)\,\mathbf{n}, \qquad \ldots\ldots(44)$$

$$\mathbf{s}' = (\mathbf{s}_1 . \mathbf{b})\,\mathbf{b}' + (h' + i\delta')\,\mathbf{n}', \qquad \ldots\ldots(45)$$

and the corresponding waves suffer an exponential decrease of amplitude with perpendicular distance from the boundary and are surface waves.

The wave normals associated with such slowness vectors have complex direction cosines and the displacement vectors also possess complex components since equation (11) now represents three complex rather than three real equations. The waves therefore have the form

$$(\mathbf{p} + i\boldsymbol{p}) \exp\left[-\omega\delta\mathbf{n} . \mathbf{r}\right] \exp i\omega[\mathbf{r} . \{(\mathbf{s}_1 . \mathbf{b})\,\mathbf{b} + h\mathbf{n}\} - t]$$

and, in general, possess (i) a displacement vector elliptically polarized in a plane of normal $\mathbf{p} \times \boldsymbol{p}$ which is oblique to the plane of incidence, (ii) exponential attenuation and sinusoidal variation with perpendicular distance from the boundary, (iii) flux of energy parallel to the boundary (Synge 1957).

The existence of surface waves at an interface makes the coefficients in the six equations for a_i and a_i' a stage further intractable from an algebraic standpoint. However, with modern computing aids, the solution to numerical cases of justifiable interest is feasible. The total flux of energy due to the several reflected and refracted waves across unit area of the boundary must balance the incident flux; this condition should serve as a check on the final results of any computation.

4.3. *Determination of Possible Reflected and Refracted Waves*

As an illustration of the problem we may consider the system shown in figure 12 which represents a plane boundary between two crystals of copper of different orientation.

The general problem is specialized and simplified to some extent by choosing the plane of incidence and reflection in the first crystal to be $(0, 0, 1)$ while the plane of refraction in the second crystal is $(1, \bar{1}, 0)$; the normal to the boundary in the first crystal, $\mathbf{n} = (\sin 10°, \cos 10°, 0)$ and in the second, $\mathbf{n}' = ([1/\sqrt{2}] \sin 4°, [1/\sqrt{2}] \sin 4°, \cos 4°)$. This means that each medium can allow the propagation of a wave whose displacement vector is normal to the paper and an incident wave of this kind (T_1) will give rise to one reflected and one refracted wave, each of the same polarization with respect to the boundary. An incident wave (L) may generate reflected L and T_2 waves and refracted L and T_2 waves. Of these, the reflected L wave will be a body wave for all angles of incidence while the refracted L wave will be a body wave until a critical angle of incidence obtains, when it will become a surface wave; of the reflected and refracted T_2 waves, each will be a body wave for all angles of incidence. Considering now the incidence of a T_2 wave: for small angles of incidence, body waves of each type will be reflected and refracted; passing through the critical angles for L waves, reflected and refracted, there is a range of incidence in which two body waves and two surface waves occur; then, as a result of the inflexions in the T_2 section of the reflecting medium, two T_2 body waves are transmitted in the first medium for a small range of incidence. In this range there occur also slowness vectors which

satisfy the boundary condition but whose associated rays are not directed away from the boundary into the appropriate medium. Finally for grazing angles of incidence, two surface waves are propagated in each medium.

In this special case, the associated rays for the incident, reflected and refracted waves all lie in the plane of incidence. In the most general form of the problem of reflection and refraction at an interface, this will not be so and only the wave normals will lie in the plane of incidence while the deviations of the rays will cause them to be oblique to this plane.

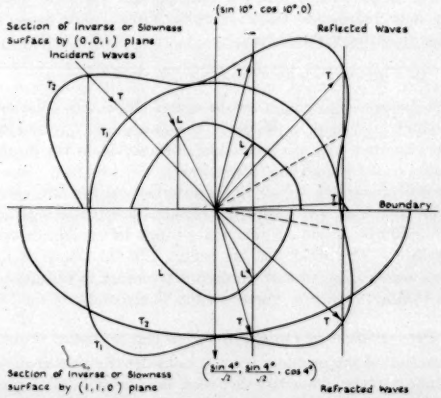


Figure 12. Reflection and refraction at a plane interface between cubic crystals of the same medium but different orientations. The possible waves are determined from the sections in the plane of incidence of the inverse or slowness surface.

§ 5. Surface Waves of Rayleigh and Love Type

5.1. *General Discussion*

In a semi-infinite isotropic medium it is possible for waves to propagate parallel to the plane boundary and to leave that surface free of stress. The Rayleigh wave possesses a composite displacement vector polarized in the plane defined by the wave normal and the normal to the boundary whose individual components suffer different exponential attenuations with perpendicular distance from the boundary; it travels with a phase velocity slightly less than that of transverse body waves in the medium. The Love wave possesses a transverse displacement vector parallel to the boundary and travels with a phase velocity slightly less than that of a transverse body wave. It can only exist when a layer of discontinuity parallel to the boundary occurs at a depth large compared with the wavelength; in the absence of such a layer it is impossible to satisfy the condition of zero stress at the boundary.

The existence of waves of the character examined in § 4.2 suggests the possibility of surface waves of composite displacement which might travel in a semi-infinite aeolotropic medium and leave the boundary free of stress. In its

most general form such a wave would be represented by

$$[a_1(\mathbf{p}_1+i\mathbf{p}_1)\exp\{-\omega\mathbf{n}.\mathbf{r}(\delta_1-ih_1)\}+a_2(\mathbf{p}_2+i\mathbf{p}_2)\exp\{-\omega\mathbf{n}.\mathbf{r}(\delta_2-ih_2)\}$$
$$+a_3(\mathbf{p}_3+i\mathbf{p}_3)\exp\{-\omega\mathbf{n}.\mathbf{r}(\delta_3-ih_3)\}]\exp i\omega(s_R-t),$$

where s_R is the component of slowness parallel to the boundary and is common to the three individual displacements which make up the composite vector and a_1, a_2, a_3, are the relative amplitudes of these displacements. If the three components of stress at the boundary are equated to zero, three simultaneous equations for a_1, a_2, a_3 are obtained. Non-zero solutions for these amplitudes are obtained when s_R has the values given by the roots of a determinantal equation. In this manner it is possible to envisage a generalized surface wave which has characteristics similar to the surface wave of § 4.2 but whose elliptically polarized displacement vector varies in a more complicated manner with distance from the free surface. However, the existence of such a wave depends upon the roots for s_R and Synge (1957) has shown that, in general, appropriate real roots of s_R do not exist and thus generalized Rayleigh waves may propagate only in certain directions lying in certain free surfaces. The particular value of velocity and the character of the displacement vector (when they exist) vary with the orientation of the free surface and the direction of propagation.

In this most general form of surface wave, the elliptically polarized displacement in the $(\mathbf{b}, \mathbf{n})$ plane of Rayleigh waves and the displacement transverse to this plane typical of Love waves are combined in a single wave. Stoneley (1955) has pointed out that the apparent fundamental difference between these types of wave motion occurs in the isotropic case and for certain axes of symmetry in crystals owing to degeneracy.

5.2. *Hexagonal and Cubic Media*

The propagation of surface waves in the basal plane of a hexagonal crystal (or isotropic plane of a transversely isotropic medium) has been investigated by Stoneley (1949). He shows that separate typical Rayleigh and Love waves may be propagated over such a surface.

In a further analysis of cubic crystals, Stoneley (1955) shows that the elastic constants of a cubic crystal must fulfil certain conditions if surface waves can be propagated over a $(1,0,0)$ plane. He discusses in detail a generalized surface wave and demonstrates the degeneracy into the usual Rayleigh and Love types of motion which takes place when propagation occurs along $(1,0,0)$ or $(1,1,0)$ axes.

In particular, his equation for the velocity of a Rayleigh wave propagated over the $(0,0,1)$ plane along the $(1,0,0)$ direction may be written

$$c_{11}(\rho v^2-c_{44})\left[\rho v^2-\left\{c_{11}-\frac{c_{12}^2}{c_{11}}\right\}\right]^2=c_{44}(\rho v^2-c_{11})(\rho v^2)^2, \qquad \ldots\ldots(46)$$

in the notation of this article. This becomes the usual Rayleigh wave equation for an isotropic medium if $c_{12}=c_{11}-2c_{44}$.

5.3. *Monoclinic and Orthorhombic Media*

Deresiewicz and Mindlin (1957) have investigated the possibility of surface waves which may travel along the digonal axis of a monoclinic crystal. They

apply their analysis to the case of an AT-cut quartz crystal and obtain a velocity of propagation $v = (0{\cdot}901c_{66}/\rho)^{1/2}$ for a typical Rayleigh wave which possesses only exponential decrease with depth.

They also treat the case of propagation along the direction $(1, 0, 0)$ over a free surface whose normal is $(0, 1, 0)$ in an orthorhombic medium and find the cubic equation for $q(=\rho v^2/c_{66})$

$$(1-d)q^3-(1-e+2f)q^2+f(f+2)q-f^2=0, \qquad \ldots\ldots(47)$$

where $d = c_{66}/c_{22}$, $e = c_{11}/c_{22}$, $f = (c_{11}c_{22}-c_{12}^2)/c_{22}c_{66}$. This equation reduces to Stoneley's equation for cubic symmetry (46) and the Rayleigh wave equation for isotropic media when the equalities in the elastic constants appropriate to the higher symmetries are inserted.

§ 6. Stationary Waves

Specimens for physical experiments are commonly of finite dimensions and, if the whole specimen is subjected to a disturbance, there are boundary conditions to satisfy. We have seen the complicated nature of reflection and refraction phenomena when an incident wave can stimulate three further waves in each of two media. When we recollect that in the simpler case of isotropic media, most problems of resonance are not soluble in closed form and in some cases remain imperfectly formulated, it should not occasion serious disappointment that the resonance of crystals of finite dimensions is far from being mathematically consolidated.

However, the very wide use of quartz and other materials for vibrational experiments and other applications has occasioned much effort directed towards an understanding, albeit partly empirical, of the vibrations of crystals. An adequate appreciation and discussion of this work would form the substance of a separate article.

Briefly, it may be pointed out that as a consequence of these difficulties the assumptions entailed in relating theoretical analysis to experimental observations are rather numerous and, which is more important, difficult to define accurately.

Notable in the considerable literature of the subject are discussions by Chree (1890), Voigt (1910), Ekstein (1945), Cady (1946), Mason (1950), Bechmann (1952) and Newman and Mindlin (1957). An extensive bibliography may be found in reviews by Hearmon (1946, 1956).

§ 7. Propagation of Elastic Waves in Polycrystalline Aggregates

7.1. *Relation of the Single Crystal to the Aggregate*

A vast amount of work involving the propagation of elastic waves in polycrystalline media has been done and a great deal of useful information has stemmed from it. Almost always, an inherent assumption about such media is that they are isotropic; that is, the individual molecules or grains, small compared with the wavelength, each possessing inherent aeolotropy, are aggregated with a random distribution of their individual orientations.

This assumption of isotropy can be justified in very many cases but it is well known that oriented polymer materials, plastically deformed metals, layered or

sandwich structures, etc. are fundamentally anisotropic and it is clearly desirable to know which physical properties may still be isotropic.

The physical properties, such as the elastic constants, of aggregates are in some way related to the corresponding properties of the fundamental unit but the precise form of the relation remains obscure. Many years ago, Voigt (1910) devised an averaging procedure which, when applied to the elastic constants of a crystal, yielded the effective constants of an isotropic aggregate which was experiencing uniform strain. Later, Reuss (1929) presented a similar procedure which yielded the effective constants for an aggregate experiencing uniform stress. These average values are discrepant by an amount which is dependent on the inherent aeolotropy of the unit crystals and, in general, they bound the observed values. More recent discussions include those of Bruggeman (1934) and Hill (1952).

However, the formidable problem of relating the unit to the aggregate remains a challenge and the sensitivity of velocity measurements may provide very useful information towards a solution.

7.2. *Elastic Anisotropy due to Preferred Orientation*

Preferred orientation of the units of an aggregate may cause elastic anisotropy. This will be reflected in variation of velocity with orientation of the wave normal in measurements on progressive waves or anomalies in harmonic series of resonant frequencies in measurements on stationary waves.

Serious preferred orientation, especially of units which are inherently highly aeolotropic, may cause significant variation of velocity inevitably accompanied by extraordinary refraction. Such effects could prove embarrassing, if not worse, in the technology of flaw detection while imperfectly understood.

Measurements of the resonant frequencies of rolled cylindrical bars have been found anomalous and a method of interpreting such results has been given by Pursey and Cox (1954). Their analysis relates to the particular case of aggregates of cubic crystals which, owing to working, possess apparent hexagonal symmetry and leads to one result of great interest for which there is some supporting evidence. They express the preferential orientation as a probability function in terms of spherical harmonics and find that only the fourth order harmonic affects the behaviour of elastic waves. It is apparent that this result is dependent upon the use of a simplifying assumption about the variation of velocity with orientation; however, a number of elastic observations have been made on specimens which x-ray analysis shows to be extremely preferentially oriented and little elastic anisotropy has been found. There is some evidence also that aggregates of face-centred cubic crystals are more susceptible to elastic anisotropy as a result of working than aggregates of body-centred cubic crystals. It would seem therefore that such induced elastic anisotropy is significantly dependent upon the slip system of the basic unit.

A substantial investigation of these effects would, I believe, be fruitful.

7.3. *Elastic Anisotropy due to Laminated Structure*

Composite media of different isotropic layers are often known as transversely isotropic. When the thickness of individual layers is small compared with wavelength, many of the characteristics of elastic wave propagation are similar to those which obtain in hexagonal media.

It is this type of anisotropy which is most significant in seismological discussions.

A further interesting example is afforded by coal which, according to the suggestions of Terry and Morgans (1958), possesses elastic anisotropy largely as a result of small cracks between the layers of its structure. When compressed normal to the layers so that the cracks tend to close, a marked diminution of elastic anisotropy occurs.

7.4. *Problems of Attenuation*

In this essay, we have discussed the propagation of mechanical waves in truly elastic crystals ; no attempt has been made to consider absorptive media. Even with this limitation, however, problems of attenuation occur in considering the behaviour of waves in aggregates. These are due to the scattering of waves caused by the mismatch in mechanical impedance which occurs at boundaries. The principal factors which affect this type of attenuation are the ratio of the wavelength propagated to the size of unit in the aggregate and the magnitude of the inherent aeolotropy of the unit. There has been some investigation of this type of attenuation by Roth (1948), Mason and McSkimin (1948), Lifshitz and Parkhomowski (1950) ; the last paper appears to deal with the problem in the most satisfactory manner but its value is severely limited by the apparent inaccessibility of a previous paper to which reference is desirable.

§ 8. Relation between Theory and Experiment

It is appropriate to end this review with some consideration of how closely the theoretical and practical aspects of this subject correspond and an indication of the assumptions which are necessary in relating the two.

Strictly, the analysis which has been sketched here deals with continuous infinite plane waves travelling in an aeolotropic Hookean elastic continuum ; the progressive envelope of a set of such wavefronts having a common point of intersection at a given instant has also been found. What are the assumptions it is necessary to make about experimental conditions in order that observations may be interpreted in terms of these theoretical concepts ?

In order that

(*a*) a crystal lattice may be considered a continuum, the elastic wavelength propagated must be much greater than the lattice spacing :

(*b*) Huygens' principle and a ray theory may be appropriate, the elastic wavelength propagated must be much smaller than the dimensions of the specimen and of any effective aperture through which the waves pass.

These two conditions are here expressed qualitatively but they may be given quantitative interpretation to any desired accuracy provided the terms they contain have precise meaning. In this respect, some further consideration of elastic wavelength is required when pulses or wave packets of mechanical energy are used in experiments since the existence of pulses implies an infinite frequency spectrum.

In the absence of any dispersive effects, detectable by an alteration in the profile of the pulse, it seems reasonable to argue that only the amplitudes associated with the frequencies of the waves within the packet and of the pulse repetition are significant. The latter frequency is very much lower than the amplifier, which

assists in the display of the pulse, can affect ; consequently, measurements may be related to the former frequency which presumably is central in a band of small width. It is on some basis such as this that variations of elastic constants with frequency have been reported. However, if dispersive effects become evident it is clear that the behaviour of the pulse is dependent upon differential effects over a wide band of frequencies and no single frequency may justifiably be associated with measurements on such pulses. This difficulty calls into question a large number of results on damping coefficients and their variation with frequency. There appears to be an opportunity for a critical examination of the physical meaning of such results.

Finally, it would be fair to claim that the propagation of elastic waves in any given crystal can now be formulated and understood in a quantitative manner although heavy computation may be involved. The propagation of mechanical waves in absorptive crystals presumably follows a similar pattern to the behaviour of light waves under analogous conditions. Further understanding of the relation between unit and aggregate would have practical advantages for the metallurgist and engineer besides affording philosophical interest and satisfaction. It is hard to believe that a search on a broad front for such understanding would be an unprofitable venture in scientific and industrial research.

Acknowledgments

I wish to thank Dr. K. Helbig, of the Institute for Applied Geophysics, Munich, for making available the photograph from which figure 11 has been prepared.

My thanks are also due to Mr. H. L. Cox and Mr. R. F. S. Hearmon, both of whom read the article in draft and offered valuable criticism.

The line drawings are reproduced by permission of the Director, National Physical Laboratory.

References

Alexandrov, K. S., 1956, *Kristallographia*, **1**, 718 ; 1957, *Ibid.*, **2**, 140.
Alexandrov, K. S., and Khanmov-Mal'kov, V. Y., 1956, *Kristallographia*, **1**, 373.
Bechmann, R., 1952, *Proc. Phys. Soc.* B, **65**, 368 ; see also articles in *Piezoelectricity*, 1957 (London : H.M.S.O.).
Blanchet, P. H., 1840, *Journal de Mathematiques* (Lionville), **5**, 1 ; **7**, 13.
Borgnis, F. E., 1955, *Phys. Rev.*, **98**, 1000.
Bruggeman, D. A. G., 1934, *Z. Phys.*, **92**, 561.
Cady, W. G., 1946, *Piezo-electricity* (New York : McGraw-Hill).
Cauchy, A., 1836, *C.R. Acad. Sci., Paris*, **2**, various references in this volume.
Chadwick, P., and Sneddon, I. N., 1958, *J. Mech. Phys. Solids*, **6**, 223.
Chree, C., 1890, *Quart. J. Math.*, **24**, 340.
Christoffel, E. B., 1877, *Ann. di Mat.*, **8**, 193.
Deresiewicz, H., 1957, *J. Acoust. Soc. Amer.*, **29**, 204.
Deresiewicz, H., and Mindlin, R. D., 1957, *J. Appl. Phys.*, **28**, 669.
Ekstein, H., 1945, *Phys. Rev.*, **68**, 11.
Fresnel, A., 1827, *Oeuvres* (Paris : Imprimerie Imperiale).
Green, G., 1842, *Trans. Camb. Phil. Soc.*, **7**, 121.
Hamilton, Sir W. R., 1832, see 1931, *Collected Papers* (Cambridge : University Press).
Hearmon, R. F. S., 1946, *Rev. Mod. Phys.*, **18**, 409 ; 1956, *Phil. Mag. Suppl.*, **5**, 323.
Helbig, K., 1956, *Geophysical Prospecting*, **4**, 70.
Hill, R., 1952, *Proc. Phys. Soc.* B, **65**, 349.

HUYGENS, C., 1690, see translation 1912, *Treatise on Light* (translated by Silvanus P. Thompson) (London : Macmillan).
JOEL, N., and WOOSTER, W. A., 1957, *Nature, Lond.*, **180**, 430.
KELVIN, LORD, 1904, *Baltimore Lectures* (Cambridge : University Press).
DE KLERK, J., and MUSGRAVE, M. J. P., 1955, *Proc. Phys. Soc.* B, **68**, 81.
LAVAL, J., 1951, *Institut Internationale Physique Solvay Conseil*, Brussels.
LECORRE, Y., 1953, *C.R. Acad. Sci., Paris*, **236**, 1903.
LIFSHITZ, I. M., and PARKHOMOWSKI, G. D., 1950, *Zh. Eksper. Teor. Fiz.*, **20**, 175.
MAKINSON, K. R., 1955, *Proc. International Wool Textile Research Conference, Australia* (Melbourne : C.S.I.R.O.), Pt. D, p. 54.
MARKHAM, M. F., 1957, Conference on Physics of Non-destructive Testing, *Brit. J. Appl. Phys.*, Suppl. No. 6, S.56.
MASON, W. P., 1950, *Piezo-electric Crystals and their Application to Ultrasonics* (New York : van Nostrand).
MASON, W. P., and MCSKIMIN, H. J., 1948, *J. Appl. Phys.*, **19**, 941.
MCCULLAGH, J., 1848, *Trans. Roy. Irish Acad.*, **21**, 17.
MILLER, G. F., and MUSGRAVE, M. J. P., 1956, *Proc. Roy. Soc.*, **236**, 352.
MUSGRAVE, M. J. P., 1954 a, *Proc. Roy. Soc.*, **226**, 339 ; 1954 b, *Ibid.*, **226**, 356 ; 1957 a, *Acta Cryst., Camb.*, **10**, 316 ; 1957 b, *Proc. Camb. Phil. Soc.*, **53**, 897.
NEIGHBOURS, J. R., BRATTEN, F. W., and SMITH, C. S., 1952, *J. Appl. Phys.*, **23**, 389.
NEWMAN, E. G., and MINDLIN, R. D., 1957, *J. Acoust. Soc. Amer.*, **29**, 1206.
NEWTON, SIR ISAAC, 1704, see translation 1931, *Opticks*, Bk. 3 (London : Bell).
PURSEY, H., and COX, H. L., 1954, *Phil. Mag.*, **45**, 295.
RAMAN, SIR C. V., and VISWANATHAN, K. S., 1955, *Proc. Indian Acad. Sci.* A, **42**, 1, 51.
REUSS, A., 1929, *Z. angew. Math. Mech.*, **9**, 49.
ROTH, W., 1948, *J. Appl. Phys.*, **22**, 1089.
RUDSKI, M. P., 1911, *Anz. Akad. Wiss. Krakau Bull.*, **8a**, 503.
STONELEY, R., 1949, *Mon. Not. Roy. Astr. Soc.* (Geophysical Suppl.), **5**, 343 ; 1955, *Proc. Roy. Soc.* A, **232**, 447.
SYNGE, J. L., 1956, *Proc. Roy. Irish Acad.*, **58**, 13 ; 1957, *J. Math. Phys.*, **35**, 323.
TERRY, N. B., and MORGANS, W. T. A., 1958, *Conference on Mechanical Properties of Non-metallic Brittle Materials* (London : Butterworths Scientific Publications).
VOIGT, W., 1910, *Lehrbuch der Kristallphysik* (Leipzig : Teubner).
WHITTAKER, SIR EDMUND, 1951, *History of the Theories of the Aether and Electricity* (London : Nelson).

RELATIVITY

By H. BONDI
University of London, King's College

CONTENTS

PAGE

PART I: SPECIAL RELATIVITY

§ 1. Newtonian relativity ... 97
§ 2. The special principle of relativity ... 99
§ 3. Space–time ... 101
§ 4. Relativistic dynamics ... 103
§ 5. The status of special relativity ... 105

PART II: GENERAL RELATIVITY

§ 1. Newtonian gravitation ... 106
§ 2. The principle of equivalence ... 107
§ 3. Riemannian formulation ... 108
§ 4. The general theory of relativity ... 110
§ 5. Modern developments ... 113
5.1. Conservation laws ... 113
5.2. Equations of motion ... 114
5.3. Gravitational waves ... 116
5.4. The Hamiltonian formulation and quantization ... 117
5.5. Cosmology ... 117
5.6. Electromagnetism ... 118

Acknowledgments ... 119

References ... 119

Abstract. The laws of relativity corresponding to Newtonian dynamics are discussed and contrasted with those suited to Maxwell's equations. Einstein's principle of relativity is inferred, and some of its consequences are briefly examined.

The peculiar features of Newtonian gravitational theory are described and are shown to lead to the principle of equivalence. The roles of this principle and of the general principle of relativity in the foundations of Einstein's general theory of relativity are critically discussed, and the theory is formulated. Modern developments in the theory are summarized.

I. SPECIAL RELATIVITY

§ 1. NEWTONIAN RELATIVITY

IT is the purpose of every physical theory to describe in a concise manner a wide variety of phenomena. In many cases this will necessitate, as part of the theory, a prescription for applying the theory to systems in differing states of motion. A prescription of this kind, a code of translation as it were, will generally consist of a mathematical system of transformation laws.

It is in the nature of transformation laws to change most quantities but to leave some quantities unchanged. These are called the invariants of the transformation and serve to define its character. A physical statement of what these invariants are is called a principle of relativity, and the fundamental equations of a theory usually define the principle of relativity applicable to it.

Newtonian dynamics furnishes a particularly important and instructive example. According to Newton's first law, a body sufficiently far removed from all other bodies will have a constant velocity vector. Since velocity is a relative quantity, it follows that, to give meaning to Newton's first law, it is necessary to specify how velocity is to be measured. For this purpose we introduce a concept of great utility in the whole subject of relativity, the observer. The basis of this concept is the simple idea that physics is the result of the labours of physicists, that for biological, sociological and technological reasons it happens that all physicists known to us are always on or within a few miles of the surface of the Earth, are working now or have been doing so during the last few centuries (or at most millenia), and have velocities relative to the Earth that are less than 1% of the velocity of light. However, the reasons for these limitations are all extraneous to physics, and so we imagine the universe peopled with physicists free of these limitations. These imaginary scientists are called *observers* and form a useful extension of the class of actual physicists.

If we now examine the significance of Newton's first law against this background of observers we find that the law will appear to be correct to some but not to all observers. For the restriction to bodies sufficiently far from all other bodies (though unclear cosmologically and possibly even philosophically) is effectively independent of the observer, whereas the constancy of the velocity is clearly dependent on the observer measuring it.

Let us now take an observer who moves in such a way that Newton's first law appears correct to him. Such an observer may be pictured as being himself attached to a particle moving with constant velocity vector. Then with a minimum of assumptions concerning the transformation laws for distance and time (homogeneity of space and time, isotropy of space, additivity of collinear distances and of time intervals) it follows that Newton's first law will also appear to be correct to any observer moving relative to the first one with constant velocity vector.† On the other hand, to an observer accelerated relative to the first one, Newton's first law will appear to be incorrect. This is easily seen for if this observer for a time sits on a particle moving with constant velocity and then accelerates away from it, he will judge its velocity relative to him to have been zero at first and then to have increased.

It follows therefore that Newton's first law divides all observers into two classes, viz. those who find the law to be correct (called inertial observers) and others. *The inertial observers form a family in the sense that each one of them is fully equivalent to (and indeed by dynamical experiment indistinguishable from) any other inertial observer.* This statement may be called the Newtonian Principle of Relativity. Any inertial observer sees every other inertial observer as moving with constant velocity vector, whereas he sees every non-inertial observer following an accelerated motion.

This distinction is vital and persists throughout the whole subject. Why there should be this preferred family of inertial observers is a question of great obscurity which will be briefly discussed later as the problem of Mach's principle. It should be noticed that the primitive notion of an absolute space and time,

† Every inertial observer is supposed to use a rigid non-rotating system of coordinates extending throughout space.

while consistent with a set of inertial observers, does not agree with the spirit of the theory, as all inertial observers are fully equivalent (i.e. there is no sense in which one is preferred to all the others). The state of rest in this hypothetical absolute space is therefore observationally indistinguishable from that of uniform motion.

Newton's second law completely specifies the laws of transformation between any two inertial observers. For the second law specifies that *force* and *mass*, which are considered physically real and independently measurable quantities, are linked to *acceleration*. Accordingly, if any inertial observer measures the acceleration of a body, his result must agree with that of any other inertial observer. Therefore acceleration is an invariant of the transformation connecting the measurements of any two inertial observers. It follows that if $\mathbf{r}, \mathbf{r}'$ are the position vectors of an event as measured by two inertial observers moving with relative velocity $\mathbf{v}$ and t, t' are the corresponding time measurements, then

$$\left.\begin{aligned} \mathbf{r}' &= \mathbf{r} - \mathbf{v}t + \text{const.}, \\ t' &= t + \text{const.}, \end{aligned}\right\} \quad \ldots\ldots\text{(I.1)}$$

where the constants refer only to the choice of origin. These are known as Galilean transformations and are thus the transformation rules of Newtonian dynamics.†

It is essential to realize at this stage the fundamental unity of physics. There is probably no experiment involving Newtonian dynamics in which electromagnetic phenomena (including, as they do, the existence of rigid bodies and of optical means of observation) do not play an essential part. This fundamental unity demands that the transformation laws should be the same for all parts of physics, for otherwise different inertial observers would obtain essentially different analyses of the same processes.

§2. The Special Principle of Relativity

It may well be that if Newtonian dynamics had not been used so successfully for so long a time before Maxwell formulated the laws of electromagnetism, the ensuing difficulty would have been resolved more easily. For while Maxwell's equations distinguish, just like Newton's first law, between one family of mutually unaccelerated observers and all others, they are not invariant under the Galilean transformations (I.1) but under a different set of transformations. That this must be so is fairly obvious for while Newton's laws only contain acceleration, not velocity, Maxwell's equations refer directly to a special velocity which indeed is of cardinal importance, the velocity of light. The Galilean transformations (I.1) do not leave any velocity unchanged and hence their application to Maxwell's equations leads to a set of equations differing from Maxwell's, implying amongst other consequences that light travels with different velocities in different directions.

The basic conflict can therefore be put into the following terms: Newton's first law and equally Maxwell's equations single out a preferred family of observers, the inertial observers. If Galilean transformations govern the relation between

† The rules for transformation to non-inertial observers (' moving axes ') can be obtained easily, and show that such an observer would find as the cause of motion of bodies not only real forces but also fictitious ones (centrifugal force, etc.).

two such inertial observers, then acceleration (and with it Newton's second law) is an invariant, but the velocity of light (and hence the form of Maxwell's equations) is not an invariant. Accordingly, while any two inertial observers would be dynamically equivalent (i.e. local dynamical experiments by each would give the same result) they would not be electromagnetically equivalent, and so local electromagnetic experiments could discriminate between different inertial observers and could in particular determine a state of absolute rest. Such a state of affairs would be utterly contrary to the essential unity of physics, and so it is most gratifying that a series of experiments shows clearly that this difficulty does not occur. The most celebrated is the experiment of Michelson and Morley, which shows that, in spite of the Earth's velocity in its motion round the Sun, the propagation of light on the Earth takes place according to Maxwell's equations.

The resolution of the difficulty is to abandon the Galilean transformations (I.1) and to use instead the transformations that leave Maxwell's equations unchanged. Then the difficulty arises that now acceleration is not an invariant and so Newton's second law will not be unchanged under the transformation. This difficulty, however, can be overcome by so modifying Newton's second law that it too is invariant under this new transformation. That this modification is possible without conflicting with the wealth of dynamical evidence which was known to corroborate Newton's second law at the turn of the century is due to the fact that while Maxwell's equations and so their characteristic transformations are concerned with the velocity of light, classical mechanics deals only with very much smaller velocities. All that is required, therefore, is that these new transformations should have the Galilean transformation (I.1) as a limit when $|\mathbf{v}|$ is very much smaller than c, the velocity of light, and this is indeed the case.

The result of all this is simply a restatement of the Newtonian Principle of Relativity which now, however, also includes electromagnetism. Einstein's Principle of Relativity therefore states that all inertial observers are fully equivalent in the sense that if any one of them carries out physical experiments he will always obtain results identical to those found by any other inertial observer carrying out the corresponding experiments.

In this form the principle refers to all physics, not merely to dynamics and electromagnetism. According to present knowledge this means that nuclear forces too are governed by the same transformation laws as the others, and the evidence in favour of this will be discussed later. The principle as stated here refers only to inertial observers and is hence called the Special or Restricted Principle of Relativity.

The mathematical content of the principle depends on the assumption that Maxwell's equations are correct. The characteristic transformation that turns Maxwell's equations into themselves in going from one inertial observer to another with relative velocity v in the x direction are the Lorentz transformations

$$\left.\begin{aligned} x' &= \frac{x - vt}{(1 - v^2/c^2)^{1/2}}, \quad y' = y, \\ z' &= z, \quad t' = \frac{t - vx/c^2}{(1 - v^2/c^2)^{1/2}}. \end{aligned}\right\} \qquad \text{.....(I.2)}$$

The revolutionary impact of Einstein's theory of relativity is largely due to the last of these equations which shows that time is not an invariant of the transformation as it was in (I.1). In particular it follows that if two events have the same values of t, but differing values of x, then their t' values will be different. In other words, two events that appear to be simultaneous (though spatially separated) to one inertial observer will not appear simultaneous to another observer. This relativity of simultaneity is a striking feature of the theory which was first clearly perceived by Einstein.

It is easily seen that the expression

$$t^2-\frac{x^2}{c^2}=t'^2-\frac{x'^2}{c^2}$$

is unchanged by the transformation. More generally, if there are two events (t_1, x_1, y_1, z_1), (t_2, x_2, y_2, z_2), then (even when (I.2) is generalized by rotations so that the relative velocity is no longer in the x direction)

$$(t_2-t_1)^2-\frac{1}{c^2}[(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2] \qquad \ldots\ldots(I.3)$$

is unchanged by the Lorentz transformation. A particularly simple case occurs when this invariant vanishes. This means that the spatial distance of the events equals the product of c and their time difference, i.e. it means that a flash of light emitted by the earlier event would illuminate the later one. Clearly all inertial observers must agree on this, as the universal validity of Maxwell's equations implies that they all see c as the velocity of light.

§ 3. Space–Time

The significance of all these formulae becomes much clearer in terms of Minkowski's four-dimensional representation. To clarify the meaning of this term one only has to recall that in ordinary three-dimensional space the position of a point is determined by three coordinates, that by adopting a differently oriented system of coordinates these will take differing values, each being made up of a combination of the old ones, and that corresponding to any two points an invariant (their distance) can be constructed from their coordinates. Just as the three coordinates of ordinary space become mixed up on transforming to a differently oriented system of coordinates, so the four coordinates (t, x, y, z) of an event become mixed up on transforming to a differently moving system. Just as an invariant of the coordinates of two points in three-space may be constructed by *adding* the squares of the differences of the three coordinates, so in four-space an invariant may be constructed by prescription (I.3), which equally involves the squares of the differences of the coordinates but with one positive sign and three negative ones. Whereas Euclidean space has a positive definite metric (i.e. the invariant, distance, is always positive), Minkowski space has an indefinite metric, the invariant (I.3) being able to take positive, zero or negative values. The invariant is denoted by s^2, the so-called interval s being therefore real, zero or imaginary. These three cases deserve separate treatment:

(i) $s = 0$. Given (t_1, x_1, y_1, z_1), the locus of (t_2, x_2, y_2, z_2) fills a three-dimensional entity in four-space which is known as the light cone. If one space dimension is

suppressed the conical character becomes quite clear (figure 1). The event (t_1, x_1, y_1, z_1) is at the vertex of the cone and divides it into two separate halves, the past light cone ($t_2 < t_1$) and the future light cone ($t_2 > t_1$).

(ii) s real. In this case (t_2, x_2, y_2, z_2) is *within* the light cone which has t_1, x_1, y_1, z_1, as its vertex, and the reciprocal relation is similarly satisfied. A unique straight line may be drawn through the two events inclined less than the light cone to the time axis and thus representing an inertial observer (straightness being equivalent to constant velocity) moving at less than the velocity of light. If we

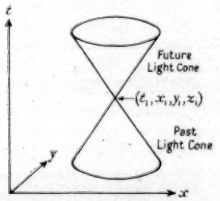


Figure 1. The light cone.

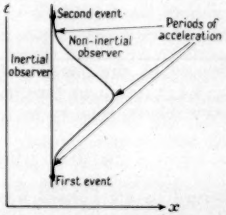


Figure 2. The 'clock paradox'.

transform to this observer, then he will assign the same (in fact zero) space co-ordinates to both events, and thus s just equals the time difference as measured by him. (s being an invariant, it is the same for him as for any other inertial observer.) A real s, therefore, is the time difference between two events as measured by the unique inertial observer passing through both events. Therefore s is called the proper time.

(iii) s imaginary. In this case the two events are outside each other's light cones. It is easily seen from (I.2), combined with a pair of space rotations, that we can find an inertial observer who will see the two events as *simultaneous*. Hence $-c^2 s^2$ equals the square of the spatial distance so that $c|s|$ is the distance between the two events as measured by an inertial observer who sees both events as being simultaneous.

A characteristic feature of all this is that while the differential dt is always a perfect differential (being simply the differential of the coordinate t), it is an invariant only in Newtonian theory (by (I.1)) but not in relativity (by (I.2)). The differential ds, while an invariant in relativity, is, by its definition

$$ds^2 = dt^2 - \frac{1}{c^2}(dx^2 + dy^2 + dz^2), \qquad \ldots\ldots\text{(I.4)}$$

equally clearly not a perfect differential, just as the differential of distance is not a perfect differential in Euclidean geometry. Thus proper time is a route-dependent quantity, whereas Newtonian time is not. This is the origin of the so-called clock 'paradox', according to which the proper time between two events measured by an inertial observer is greater than the proper time between the same two events measured by a non-inertial observer, both observers passing

through both events. If the non-inertial observer has only brief bursts of acceleration (figure 2) and between them moves with velocity v relative to the inertial observer, then the proper time measured by him (i.e. by his watch) will be $(1-v^2/c^2)^{1/2}$ times the proper time measured by the inertial observer, as is easily established.

§ 4. Relativistic Dynamics

So far attention has been entirely concentrated on the kinematical features of relativity and indeed these enable one to establish such consequences as the relativistic Doppler shift and velocity addition formulae. As these are all well known we return here to the dynamical question raised earlier: How does Newton's second law have to be modified in order to make it fit in with relativistic transformation laws?

It is clear that the components of acceleration d^2x/dt^2, etc. have extremely awkward transformation properties under the Lorentz transformations (I.2), and it is similarly clear that for slow motions ds and dt are so nearly equal that one should try to rewrite Newton's second law by replacing dt by ds. For this purpose we change to tensor notation ($t = x^0, x = x^1, y = x^2, z = x^3$) and then form the velocity 4-vector† $v^\mu = dx^\mu/ds$ ($\mu = 0, 1, 2, 3$). The three space components will, for small velocities, be almost equal to the components of the customary velocity 3-vector $\mathbf{v} = (dx/dt, \ldots)$. The time component dx^0/ds is, by (I.4),

$$1 = \left(\frac{dx^0}{ds}\right)^2 - \frac{1}{c^2}\left[\left(\frac{dx^1}{ds}\right)^2 + \left(\frac{dx^2}{ds}\right)^2 + \left(\frac{dx^3}{ds}\right)^2\right] = \left(\frac{dx^0}{ds}\right)^2\left\{1 - \frac{1}{c^2}\left[\left(\frac{dx^1}{dt}\right)^2 + \left(\frac{dx^2}{dt}\right)^2 + \left(\frac{dx^3}{dt}\right)^2\right]\right\}$$

so that

$$\frac{dx^0}{ds} = \left[1 - \frac{\mathbf{v}^2}{c^2}\right]^{-1/2}. \qquad \ldots\ldots(I.5)$$

This factor, which already appears in the Lorentz transformation, is a characteristic feature of the theory. In order to obtain the momentum 4-vector the velocity 4-vector has to be multiplied by an invariant which is called the rest mass m. Thus the momentum 4-vector is

$$P^\mu = m\frac{dx^\mu}{ds}.$$

A characteristic space component may be written

$$P^1 = m\frac{dx^1}{ds} = m\frac{dx^0}{ds}\frac{dx^1}{dt} = \left[\frac{m}{(1-\mathbf{v}^2/c^2)^{1/2}}\right]\frac{dx^1}{dt}, \qquad \ldots\ldots(I.6)$$

the last factor being the customary velocity component. The expression in square brackets is often referred to as the effective mass which thus increases with velocity. The interpretation of the time component of the momentum 4-vector follows from the equation

$$c^2 P^0 = c^2 m\frac{dx^0}{ds} = c^2 m\left[1 - \frac{\mathbf{v}^2}{c^2}\right]^{-1/2} = mc^2 + \tfrac{1}{2}m\mathbf{v}^2 + \ldots. \qquad \ldots\ldots(I.7)$$

† The 4-vector character of this quantity is clear since the position vector x^μ is the prototype of a vector and ds is an invariant.

The variable part of $c^2 P^0$ is thus, for small velocities, the familiar kinetic energy. The time component of P^μ may then be interpreted as $c^{-2}\times$ energy of the particle provided that the *rest energy* mc^2 is assigned to the particle. This equivalence of mass and energy (also suggested by (I.6)) is such a vital result of the theory that further confirmation may be mentioned here. An idealized experiment due to Einstein shows that Maxwell's equations immediately imply this equivalence. For consider a long box AB (figure 3) at rest on a smooth table.

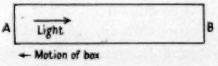


Figure 3. Mass and energy.

By the law of conservation of momentum its centre of mass must remain at rest irrespective of any internal motion. Suppose now the side A emits a flash of light towards B. The electromagnetic equations show that the emission of such a flash implies an impulse on the box due to radiation pressure equal to (energy emitted)/c. The box will therefore move in the direction BA until the flash reaches B, where it is supposed to be absorbed. Radiation pressure now exerts an equal and opposite impulse so that the box is reduced to rest, but in a different position from the original one. The centre of mass remaining in the same position, the flash of light must have transported a mass equal to its energy/c^2 from A to B. Thus again mass = energy/c^2.

The equivalence of mass and energy is most impressively demonstrated in the mass defects of the nuclei corresponding to their binding energies. This fact shows that, although the theory of relativity is based on electromagnetism, its results apply equally to nuclear forces. This is a most important extension of the theory and helps to justify the principle of relativity in the form that no *physical* experiment can discriminate between different inertial observers.

It will be seen that the velocity 4-vector $v^\mu = dx^\mu/ds$ always satisfies the relation $(v^0)^2 - c^{-2}[(v^1)^2+(v^2)^2+(v^3)^2] = 1$. In the language of the theory this means that v^μ is a unit vector. Similarly,

$$(P^0)^2 - \frac{1}{c^2}[(P^1)^2+(P^2)^2+(P^3)^2] = m^2.$$

Thus the energy $c^2 P_0$ is connected with the momentum (P^1, P^2, P^3) by

$$(c^2 P_0)^2 = m^2 c^4 + c^2[(P^1)^2+(P^2)^2+(P^3)^2]. \qquad \ldots\ldots(\text{I.8})$$

This important relation applies equally to material particles of low velocity (second term small compared with the rest mass term), of high velocity, and to immaterial particles of no rest mass (photons).

It is natural to define the force 4-vector F^μ by $F^\mu = dP^\mu/ds$. The three spatial components reduce to the Newtonian ones for sufficiently low velocity while the time component is the rate of doing work. An alternative derivation of this theorem would have been to postulate the *exact* validity of Newton's second law in the inertial frame in which the particle is instantaneously at rest. The transformation laws then give equally all the results of this section.

A simple collision problem may illustrate the working of the theory. A particle of rest mass m and velocity $\mathbf{v}$ collides elastically with another particle of the same rest mass previously at rest. What can be said about the directions of motion of the two particles after the collision?

Taking the x axis in the original direction of motion then before the impact the momentum 4-vectors of the two particles are

$$m(1-v^2/c^2)^{-1/2},\ mv(1-v^2/c^2)^{-1/2},\ 0,\ 0 \quad \text{and} \quad m,\ 0,\ 0,\ 0.$$

As energy and momentum are conserved, the sum of the two 4-vectors after the collision must equal the sum before the collision. If the directions of motion of the two particles make angles θ, ϕ with the x axis, then the conservation laws lead, through some tedious algebra, to

$$\tan\theta\tan\phi = \frac{2(1-v^2/c^2)^{1/2}}{1+(1-v^2/c^2)^{1/2}}. \qquad \ldots\ldots(\text{I.9})$$

This may be compared with the Newtonian law $\tan\theta\tan\phi = 1$, and shows that in the main the products of high velocity collisions emerge at smaller angles to the original direction of motion than would be expected on the basis of Newtonian dynamics. This result is fully confirmed by experiment.

An alternative way to derive (I.9) is to view the collision from a frame moving such that before the collision the two particles have equal and opposite velocities (the so-called centre-of-mass frame). Of course velocities have to be added in accordance with the relativistic rules which are an immediate consequence of the Lorentz transformations. By conservation of momentum the velocities of the particles after collision must still be equal and opposite, by conservation of energy they must be of the same magnitude as before, but the direction may be different. Transforming back from the centre-of-mass system to the original one, equation (I.9) is obtained.

§ 5. The Status of Special Relativity

To return to more general considerations, the crucial role of the velocity of light in the theory is clear. Since time and space are linked, a velocity is required for dimensional reasons, and this is the role of c. In converting lengths into times, does it act as an intrinsic constant of nature or is it merely a scale factor based on the units in which we happen to measure length and time?

There are strong reasons to believe that the second alternative is correct and that c is simply a conventional unit. Not only do we know that, by using radar methods, all lengths can be measured in times, but in fact the size of our measuring rods is determined by atomic interactions which themselves are fully characterized in principle by the atomic *frequencies*. Thus only time standards are true primary standards, and units of length are determined from them by the use of c, which is thus, as Milne first perceived, a purely conventional quantity. Any so-called principle of the constancy of the velocity of light is thus merely a statement of intent to use standards of length defined in a certain way.

The special theory of relativity or, as it might be called, the theory of the equivalence of inertial observers has been confirmed by countless experiments.

It is so clearly a consequence of all we know about electromagnetism and dynamics that the theory is not in serious doubt.

At the same time a curious point must be mentioned. If astronomical observations of the universe around us are made, then an appropriate anisotropy in the observed relation between the red shift of distant galaxies and their brightness would indicate a velocity of us relative to a 'rest frame of the universe' (a velocity of a few per cent of c would be readily detectable, a much smaller one could be measured in principle). The principle of relativity states effectively that there are no *locally* observable consequences of this 'absolute' velocity. While this is a correct description of all the known laboratory physics, it should not cause too much surprise if a locally observable consequence of this cosmically defined absolute velocity were discovered one day. The principle of relativity should therefore not be regarded as an absolute and unshakable truth but rather as a convenient description of a huge set of experimental data.

II. GENERAL RELATIVITY

§ 1. Newtonian Gravitation

It has been seen how Maxwell's equations lead directly to the Special Theory of Relativity, which may therefore be looked upon as part of electromagnetic theory. Indeed, as Einstein has said, all the evidence that supports Maxwell's theory thereby supports special relativity. The forces that are encountered in nature are all electromagnetic, nuclear† or gravitational in character. Special relativity is designed to deal with electromagnetic forces and, as has been pointed out, appears to deal satisfactorily with nuclear energies. What is its relation to gravitation?

In order to elucidate this question it may be advantageous to discuss first the status of gravitation in Newtonian theory. There gravitation enters as one of the possible forces, that is, as one of a set of listed quantities which may be put on the right-hand side of Newton's second law. But it is distinguished from all other forces by three remarkable characteristics:

(i) Whereas in the case of all other forces the ratio of the force on two different bodies may be varied more or less at will, in the case of gravitation this ratio always equals the ratio of their masses. As Galileo had already found, all bodies fall equally fast. In other words, gravitation determines the *acceleration* of a body without reference to its constitution, the force always being just that required to produce this acceleration. This fact is known as the equality of the inert and gravitational masses. The best experimental support is due to Eotvös, but a modern re-determination of the equality is greatly overdue.

(ii) The forces arising from any cause other than gravitation always tend to diminish their sources. Electric charges attract their opposites, nuclear forces saturate, etc. In other words, any non-gravitational field of force has positive definite energy. Gravitation is entirely different. Owing to the attraction between sources the tendency is for sources to build up, for fields to get stronger. Only the intervention of non-gravitational forces (pressures, rigidity, degeneracy) sets a limit to this build-up of gravitational forces.

† This term is used here for brevity to cover both weak and strong interactions between elementary particles.

To use an analogue, energy as the measure of exchange value plays the part of money in an economy. Everything has positive value, i.e. is bounded from below, except gravitation which thereby appears to open up unlimited credit facilities. This would destroy the utility of the whole monetary system were it not that owing to the relative weakness of gravitation the intervention of pressures, etc. usually sets an early limit to these credits. Only very occasionally does gravitation gain the upper hand (supernovae) with catastrophic results.

(iii) The absence of repulsive gravitational forces (negative mass). This is in remarkable contrast to electromagnetic forces. In some sense this fact is a consequence of (i) but this is not entirely true, and in any case depends on the absence of negative inertial masses, a similarly remarkable fact.

The General Theory of Relativity, as will be shown below, fully accounts for (i), makes reasonable provision for (ii), but is no clear advance over Newtonian theory as far as (iii) is concerned.

§ 2. The Principle of Equivalence

The first step towards general relativity is to widen the meaning of (i). Galileo's statement that all bodies fall equally fast implies that a freely falling observer sees neighbouring particles following motions without relative accelerations. In other words, an observer in a freely falling box could not discover the presence of a gravitational field by *local* dynamical experiments. The restriction to local experiments is severe and vital since all actual gravitational fields are essentially inhomogeneous, and unless 'local' is understood sufficiently strictly these inhomogeneities can be discovered by sufficiently refined experiments. As an example, consider an observer in a lift falling down a shaft extending to the centre of the Earth. Two particles at rest in the lift at the start of the motion will collide when the lift reaches the centre of the Earth, thus revealing to a sufficiently careful observer in a sufficiently wide lift the existence of a gravitational field. In the present discussion the skill of the observer is to be taken as sufficiently poor in relation to the width of the lift to exclude the possibility of such observations.

As for the restriction to *dynamical* experiments (which is due to the description of Galileo's approach) it will be clear from the discussion of the essential unity of physics given earlier that such a restriction runs counter to the spirit of the subject. More physically, the electromagnetic and nuclear nature of matter would make nonsense of the interpretation of Galileo's and Eotvös' work if these forces behaved differently. Therefore the restriction to dynamical experiments should be dropped. Finally, it is clear that if an observer is in field-free space in a box accelerated by external means he will discover in his box that all objects fall with the same acceleration (his own) owing to their inertial properties. We thus arrive at Einstein's principle of equivalence: an observer cannot discriminate by *local* experiments between inertia and gravitation.

This principle invites a re-consideration of Newton's laws. According to the first law inertia alone will make a body move with constant velocity, and this motion is taken as the *standard* against which all acceleration is measured. By this means our attention is drawn to the forces that exist in nature and make bodies deviate from the standard motion. Virtually all such forces on a body can be reduced arbitrarily by taking suitable precautions, i.e. by seeing that no rope pulls

the body, that it has no charge which can be pulled by electric fields, etc., all forces, that is, other than gravitation. Gravitation, however, cannot be eliminated by any local measures. Indeed, in the sense of Einstein's principle of equivalence, one cannot discriminate locally between gravitation, which pulls a particle off the Newtonian standard, and inertia, which keeps it on the standard. The question presents itself then of whether in these circumstances it would not be wiser to abandon the Newtonian standard and adopt instead as the standard *the motion of matter under inertia and gravitation*. In other words, since the Newtonian standard of unaccelerated motion cannot be realized by local measures owing to the fact that all bodies fall equally fast, the freely falling particle is taken as the standard, and a force is looked for only when a particle deviates from this standard. Gravitation itself is therefore no longer a force, it is one with the inertial background.

Although it would seem entirely possible to proceed straightaway to the mathematical formalism for these ideas, Einstein did in fact proceed by way of what is called the general principle of relativity, which is the statement that the laws of nature have the same *form* in all systems of coordinates, and hence for all observers, whether accelerated or not. This statement should be contrasted with the special principle of relativity which, it will be remembered, said effectively that the laws of nature were the same *in fact* for all inertial observers. Thus whereas the special principle has physical content, the general principle is void of all physical meaning and is merely a mathematical challenge, a challenge to the ingenuity of mathematicians to find the same form for the laws of nature in all systems of coordinates, however different the content of these laws might be. Given sufficiently high ingenuity on the part of the mathematician the principle places no restriction whatever on the laws.

The value, if any, of the principle may at most be placed on its heuristic use. If one does not put too much confidence in the skill of mathematicians, then one would suspect the laws of nature to be such that they are fairly easily put into the form demanded by the principle. In fact, in the foundations of the theory it is effectively demanded that the laws should be of tensor form. Curiously enough, some of the most interesting conservation laws are not of tensor form (see discussion in § 5.1).

Another way of giving content to the principle is to say that it means that there are no coordinate systems in which the laws of nature take a specially simple form, i.e. no preferred systems. Although this idea has content, and is valuable in constructing general relativity, its actual validity in the theory is in doubt (see discussion of harmonic coordinates and their importance in the work of Fock in § 5.2).

Though most writers on the subject appear to find useful guidance in the general principle of relativity, the author of this article regards it as an unnecessary encumbrance. Fock has similarly stressed that the concept of covariance is irrelevant to the distinction between special and general relativity.

§ 3. Riemannian Formulation

How can mathematical shape be given to the principle of equivalence? This principle implies that at any event there exist systems of coordinates which *locally* abolish the gravitational field. Hence special relativity will be valid locally in

these systems, i.e. they are *locally* Minkowskian. On the other hand, in general there will not be a world-wide Minkowskian system since this would imply total absence of irreducible gravitational fields. Accordingly a generalization of Minkowskian geometry is required which can be reduced to it locally anywhere, but not everywhere simultaneously. Virtually the only such geometry is Riemannian. Just as in Minkowski space, four coordinates x^0, x^1, x^2, x^3 and an invariant ds appear, but the invariant is now defined in terms of the coordinate differentials by the general quadratic form

$$ds^2 = g_{\mu\nu}\, dx^\mu\, dx^\nu \qquad \ldots\ldots\text{(II.1)}$$

where in accordance with the usual convention of the tensor calculus the repeated indices μ, ν are to be summed over. The symmetric tensor $g_{\mu\nu}$ is called the fundamental tensor. Its ten free components are arbitrary functions of the coordinates apart from conditions of differentiability and a restriction that arises as follows. A real coordinate transformation can always be found that reduces (II.1) *locally* to the Minkowskian form (using units in which $c = 1$)

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2, \qquad \ldots\ldots\text{(II.2)}$$

provided that the $g_{\mu\nu}$ are such as to give the quadratic form (II.1) the correct signature -2.

In general it is impossible to reduce (II.1) to (II.2) except locally, and so a theory based on (II.1) has the required property of being locally, but not universally, Minkowskian. Of course, if the universal validity of (II.2) is postulated, then a general transformation of coordinates leads to an expression of the form (II.1), but now the $g_{\mu\nu}$ will have the special character that allows one to regain (II.2) universally. In terms of the $g_{\mu\nu}$ the condition for this to be possible is the vanishing of the curvature tensor $R_{\mu\nu\sigma\rho}$ which, owing to its symmetry properties, has only 20 free components, each of which depends linearly on the second derivatives of the $g_{\mu\nu}$ and also more generally on the $g_{\mu\nu}$ and their first derivatives.

In special relativity, by Newton's first law, the world line of a free particle is a straight line with real ds. The obvious generalization of this concept for the metric (II.1) is a geodesic with real ds, and thus the appropriate formulation of Newton's second law in general relativity is: A particle moving under the influences of inertia and gravitation alone describes a geodesic in the space (II.1). Note that a geodesic is determined by position and velocity at an instant, just as a Newtonian orbit is.

As was pointed out above, the appropriate choice of observer can eliminate gravitational effects locally except for the difference between neighbouring orbits which is an irreducible effect of the gravitational field. The difference y^μ in the coordinates of corresponding points of neighbouring geodesics is given by the equation of geodesic variation

$$\frac{d^2 y^\mu}{ds^2} + R_{\nu\sigma\rho}{}^{\mu} \frac{dx^\sigma}{ds} \frac{dx^\rho}{ds} y^\nu = 0, \qquad \ldots\ldots\text{(II.3)}$$

which shows that the intrinsic properties of gravitational fields (the true local observables) are described by the curvature tensor $R_{\nu\sigma\rho}{}^{\mu}$. Once again it is clear that the vanishing of this quantity is equivalent to the absence of gravitation.

In Newtonian theory the gravitational field is expressed in terms of *one* potential, and a combination of its second derivatives (the Laplacian) is proportional to the density of matter (i.e. the density of sources of the field). In particular the vanishing of the Laplacian is characteristic of empty space. Owing to the way matter is arranged in space, emptiness is very common, and so this is a particularly important case. The only reasonably simple condition which can be applied in general relativity and corresponds at all to the vanishing of the Laplacian is the vanishing of the Einstein tensor

$$R_{\mu\nu} \equiv g^{\sigma\rho} R_{\mu\sigma\rho\nu} = 0 \qquad \ldots\ldots\text{(II.4)}$$

which is therefore Einstein's law of gravitation in empty space. The equation corresponding to Poisson's has to be an expression in $R_{\mu\nu}$ whose vanishing implies the vanishing of $R_{\mu\nu}$ and which is in some way related to the most characteristic property of matter as the source of the gravitational field, its conservation. The tensor $R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}$ (where $R = R_\sigma{}^\sigma$) has these properties. Thus arise the field equations of general relativity

$$R_{\mu\nu} - \tfrac{1}{2}Rg_{\mu\nu} = -kT_{\mu\nu} \qquad \ldots\ldots\text{(II.5)}$$

where k is effectively the constant of gravitation and $T_{\mu\nu}$ is a tensor characteristic of the presence of matter whose construction will be discussed later.

The equations (II.4) form a highly complicated non-linear set of partial differential equations of the second order in the ten variables $g_{\mu\nu}$ whose structure is made yet more awkward by the existence of four differential identities between them (the vanishing of the covariant divergence of the left-hand side of (II.5)). The existence of such identities need not cause surprise since equations (II.4) clearly must not permit a definite solution in terms of $g_{\mu\nu}$ for otherwise no transformation of coordinates would be possible. It is clear that the $g_{\mu\nu}$ play the role of a potential in the theory and it may justifiably be asked at this stage, when the mathematical complexity of the theory emerges, why Einstein should require ten potentials of gravitation where one was good enough for Newton. The answer to this is clear when the case of no gravitational field is considered. For then a general transformation of coordinates turns (II.2) into (II.1). Thus the effects of pure inertia in a general system of coordinates are described by ten potentials $g_{\mu\nu}$. Returning to the case of a gravitational field by means of the principle of equivalence, it follows that in general coordinates ten potentials are indeed required in order to describe gravitation. A reduction in number would only be possible if a general prescription could be given within the theory of how to find a simpler system of coordinates. Apparently this can only be done in a way (harmonic coordinates) that does not appeal to many relativists, and even those who do use these coordinates still effectively work with the ten $g_{\mu\nu}$.

§ 4. The General Theory of Relativity

Fock has recently given a searching analysis of the relation between special and general relativity. He criticizes the idea that the general principle of relativity makes a major contribution, very much as has been done in the present article, and in particular stresses that general coordinates, covariance and tensor notation can be used just as well in special as in general relativity. The *invariance*

in fact of the special principle of relativity is a consequence of the *homogeneity* of space–time implied by the universal validity of (II.2) in that theory. The chief aim of general relativity on the other hand is to determine the *inhomogeneity* of space–time by solving the field equations (II.4) or (II.5). Fock stresses the confusion due to the choice of such similar names as special and general relativity for theories that in fact differ so very widely in their fundamental concepts, but it is probably too late to change names.

As it is now clear how general relativity is based almost entirely on the principle of equivalence it may be desirable to consider again the status of this principle. Its experimental basis does not of course prove the *exact* equality of inert and gravitational mass, it only shows that the deviation from equality is less than a certain low limit. This fact in itself is amply good enough to make use of equivalence in Newtonian dynamics; if the equality is not exact it is at least a splendid approximation. In general relativity, however, the equality is taken to be exact and made the foundation of a strict logical argument. If the equality turned out to be only approximate it is difficult to see how the theory could be saved. In this sense the theory is based not so much on an experiment as on an extrapolation from the experiment.

The first and most important contact between the theory and observation is made through the solution of (II.4) for a spherically symmetric static empty space–time surrounding a non-empty region and extending to infinity, the so-called Schwarzschild metric. This solution is applied to motions in the solar system and it is found:

(i) That the motion of particles in the Schwarzschild metric are almost identical with the motion of particles in the corresponding Newtonian situation. Owing to the very different starting-points of the theories this is a striking success and means that the wealth of data supporting Newton's theory of gravitation also supports Einstein's.

(ii) There is one small difference in the motion of material particles as described by Newton's and Einstein's theories. In the case of elliptical orbits, general relativity demands a slow precession of the axes of the ellipse. The difference is most marked in the case of the planet Mercury, where the discrepancy found long before general relativity between Newtonian theory (after the perturbations due to other planets have been allowed for) and observation was satisfactorily accounted for by the new theory of gravitation. The effect is much smaller in the case of other planets, but modern work indicates that they, too, follow Einstein's rather than Newton's law of gravitation.

(iii) Since no pre-relativistic synthesis between dynamics and optics was achieved, only general relativity makes a clear statement about the deflection of light rays in a gravitational field. However, the greatest effect one could reasonably expect on a Newtonian basis is half the relativity effect. The only present way of testing this prediction is at a total eclipse of the sun if bright stars happen to be close to the limb. The observations are therefore scarce and difficult but, although some doubt has been cast on this, the evidence seems to be in fair agreement with general relativity.

(iv) This, the third of the so-called crucial tests ((ii), (iii), (iv)), requires a little more discussion. The prediction of general relativity is that if two equal

clocks (i.e. of the same construction with their time-keeping qualities compared when they are in proximity) are separated, one remaining on the Earth, while the other is put on the Sun's surface, then, viewed from anywhere, the Sun clock will be slow compared with the terrestrial one, the ratio of indicated times differing from unity approximately by the difference in Newtonian gravitational potentials divided by c^2. This test can be performed with the spectral lines of some atoms, but the agreement is very poor, probably owing to effect of local physical conditions in the Sun's atmosphere. It is important to realize, however, that this shift of spectral lines can be deduced directly from well-established experimental facts (equality of inert and gravitational masses, emission and absorption of spectral lines, equality of mass and energy), and the postulate of the non-existence of a *perpetuum mobile* of the first kind. The deduction essentially envisages an endless chain of buckets from the Earth to the Sun, one side being filled with excited atoms (or nuclei), the other with de-excited ones. De-excitation takes place when the atoms are nearest to the Sun, and systems of mirrors are arranged so as to focus the light then emitted on to the atoms close to the Earth. The excited atoms have more energy and hence more mass than the de-excited ones and therefore the Sun's gravitational pull will make the chain move. If the photons received on the Earth from the de-excitations had enough energy to excite the atoms again, a *perpetuum mobile* would have been constructed. If this is to be avoided the energy of the photons must be insufficient by exactly the amount required to make the energy account balance, and this is just the Einstein shift of the spectral lines. In the Einstein shift relativity therefore successfully incorporates what is known from other considerations. If the observations were interpreted as indicating a failure of the theory of relativity they would also imply the need for fundamental changes throughout the rest of physics. Before such observations could be so interpreted, conditions in the solar atmosphere would have to be understood very much better than at present.

Observations of the Einstein shift on other stars are also too much in doubt to be of serious interest.

The Schwarzschild metric furnishes the only direct link between general relativity and observation. However, there is a great body of theory beyond this point constructed for the sake of completeness and in order to elucidate the links between general relativity and other parts of physics. The first necessity is to give content to equation (II.5) by defining the tensor $T_{\mu\nu}$ which acts as the source of the gravitational field. As there are ten potentials $g_{\mu\nu}$ in the field, there should be ten source functions forming a symmetrical tensor $T_{\mu\nu}$. In Newtonian physics it is a characteristic feature of the gravitational field that its sources are conserved (this is a much stronger condition than the corresponding one in electromagnetism since all sources have the same sign). The tensor which forms the left-hand side of (II.5) is closely related to (II.4) and has vanishing divergence. The existence of such a geometrical tensor suggests that we should look for a tensor representing mass, energy and momentum which shows conservation properties. However, conservation in Riemannian space is no simple matter. A conservation law means that the change of the integral of a quantity through the interior of a volume is related to the integral of another quantity over the surface of the volume. In Riemannian geometry the addition of tensors at different points does not result in

a tensor, and thus there are no tensor conservation laws. One can have instead: (i) a tensor which is strictly conserved whenever the Riemannian space is flat but whose behaviour in curved regions is much more complicated; (ii) strict conservation laws for non-tensorial quantities, i.e. expressions which are not transformed simply when a change of coordinates is made and often can be made to vanish locally by a suitable choice of the system of coordinates; the physical interpretation of such quantities may therefore be difficult.

Following alternative (i), $T_{\mu\nu}$ is constructed as a tensor which is strictly conserved in the absence of gravitation. It is therefore very close to the Newtonian energy concept which also, as has been seen, has a tendency to become less tangible when gravitation is included. Accordingly in matter without random motions (cold matter) the definition chosen is

$$T^{\mu\nu} = \rho v^{\mu} v^{\nu} \qquad \ldots\ldots\text{(II.6)}$$

where ρ is an invariant characterizing the density of the matter and v^{μ} is its velocity 4-vector. If there are random motions local summations are carried out leading to stress terms in the space–space parts of $T^{\mu\nu}$. The four equations $\partial T^{\mu\nu}/\partial x^{\nu} = 0$ (in flat space) are then the Eulerian equations of motion and the equation of continuity for a fluid, i.e. the equations of conservation of matter, energy and momentum in special relativity. In the presence of gravitation equation (II.5) implies that $\partial T^{\mu\nu}/\partial x^{\nu} \neq 0$ since the gravitational field transfers energy and momentum to the matter, and in the first approximation these transfer terms are the Newtonian ones. Finally the coefficient k in equation (II.5) (Einstein's gravitational constant) is found by comparison with the Newtonian case.† Thus (II.5) is supported, showing how differently gravitational and other energies are treated in relativity, which, from the point of view explained at the beginning of this section, is distinctly an advantage.

§ 5. Modern Developments

A brief survey of modern work forms the final part of this article. Naturally the author's interests and views are bound to colour strongly such a survey.

5.1. *Conservation Laws*

This is work on non-tensorial quantities that are strictly conserved in Riemannian space, and on their physical interpretation. The first of these was the so-called pseudo tensor of the gravitational field, a quantity involving the $g_{\mu\nu}$ and their first derivatives which when added to $T^{\mu\nu}$ yield a quantity which is conserved. The discovery by Landau and Lifshitz, Bergmann, Goldberg and others of a much wider class of conserved quantities has greatly deepened our understanding of the structure of the equations of general relativity and furnishes a more satisfactory basis for other work (§§ 5.2, 5.4). In particular, it has been shown that there are 'strong' conservation laws which hold whether (II.4) or (II.5) apply, and 'weak' conservation laws which depend on the field equations (II.4). The relation of these laws to the invariance properties of the theory has in particular been elucidated by Bergmann and his school.

† The need for a gravitational constant only arises because usually the density ρ in (II.6) is not measured in gravitational, but in arbitrary units (grammes/cm^3).

It should be said here that conservation laws are almost an essential for any physically useful macroscopic theory. For only conservation laws enable one to make statements about systems as a whole without detailed knowledge of their contents. A classical theory without conservation laws is thus an impossibility as far as matter is concerned, for in such a theory every result would depend explicitly on atomic or other quantal properties.

5.2. *Equations of Motion*

A short digression on negative mass in general relativity may be helpful. In the Schwarzschild solution of (II.4) the mass of the source enters as a constant of integration. If this is taken to be positive, approximately Newtonian orbits corresponding to an attractive source result. Similarly, with a negative constant the orbits are appropriate to a repulsive source. Note that these are the orbits of *all* particles (by the principle of equivalence), whatever their mass, whether positive, zero or negative. Imagine now two bodies of opposite masses at rest side by side. The positive mass attracts *all* masses and so attracts the negative one. The negative one repels *all* masses and so repels the positive one. Thus the two bodies will begin to move in the same direction. If the two masses are of suitable size they will keep their distance the same and thus the two will continue to accelerate each other and will rush off together.

In the theory of relativity mass may be described either by the energy momentum tensor $T^{\mu\nu}$ as in (II.5), or as singularities of the field, with (II.4) applying everywhere except at the singularity itself. If the first description is adopted, then it is immediately obvious that the field equations (II.5) imply the equations of motion of matter. For in Minkowski space, as has been pointed out, the equations describing the vanishing of the divergence of $T^{\mu\nu}$ are the Eulerian equations of motion, and the same is true, with the gravitational terms added, in the general case. Moreover, if only pressure-free matter is concerned, i.e. if $T^{\mu\nu}$ is of the form (II.6), then the vanishing of the divergence of (II.6) implies that elements of matter describe geodesics. The postulate that particles under gravitation and inertia only describe geodesics is therefore unnecessary, as it is a consequence of the field equation (II.5). If there are stress terms, equations of state are required to link the different components of (II.5). For, while in (II.6) there are only four independent quantities, there are ten in (II.5).

It is a remarkable tribute to the power of general relativity that the field equations (II.5) imply the equations of motion. In no other of the great field theories is this the case, and it involves a distinct economy of hypotheses. It does, though, in general involve the need to specify some equations of state of matter linking the different components of $T^{\mu\nu}$ (e.g. pressure and density).

In spite of the great success of field equations (II.5) in this direction, there have also been considerable efforts to deduce equations of motion from (II.4) through the alternative description of matter as singularities. The interest in this has been due partly to an objection to the *ad hoc* character of the construction of $T^{\mu\nu}$, partly to the need to specify equations of state, and finally to the recognition that actual matter is apparently constructed of point particles. Whether this last fact (which is essentially quantal in character) fits at all into the concept of as classical a theory as general relativity is open to debate.

Although the possibility that equations (II.4) might specify the equations of motion of point particles was recognized early, the first serious attacks on the problem were made in the late 1930's by Einstein, Infeld and Hoffman (EIH) on the one hand, and by Fock and his school on the other. The EIH method (since greatly improved and elaborated, mainly by Infeld) proceeds by a method of successive approximations and the construction of surface integrals surrounding the singularity. The nature of these surface integrals has been much clarified by the work of Bergmann, Goldberg and others on 'weak' and 'strong' conservation laws (referred to above). The successive approximations lead to very heavy calculations (substantially shortened by Infeld's work) and leave the question of convergence entirely unsolved. The vanishing of the divergence of $T^{\mu\nu}$ and the non-linearity of the equations imply that there is an interaction between the field of the singularity and the rest of the field, and this interaction defines the motion. The specification of what the motion actually is like is complicated by the existence of the singularity itself.† The use of surface integrals more or less circumvents this difficulty. However, the specification of the nature of the singularity is essential. For, as the remarks about negative mass indicated, a dipole is self-propelled. If there is any dipole or higher contribution to the singularity it will move quite differently from a simple pole. 'Real' particles are probably best represented by pure poles, but there is no clear way of specifying such a simple pole embedded in a general space. The specification has to be done step by step in the EIH method. As Bergmann has shown, the description of a singularity may require an infinity of parameters. It may well be that an equation of state for continuous matter is a lesser evil.

The approach of Fock to the problem of the equations of motion is quite different and of great importance. In general relativity one usually imposes no conditions (other than continuity and differentiability) on the coordinates. This is done because until the nature of the space is known it is not clear which coordinate systems are admissible. In flat space (special relativity) one usually works with Cartesians (which are determinate except for general Lorentz transformations), but these coordinates are usually not admissible in curved spaces. However, in dealing with specific problems one often imposes coordinate conditions because otherwise the equations obtained would not have sufficiently definite solutions. In the EIH method *ad hoc* coordinate conditions are imposed at each stage of approximation, and, as Infeld has shown, a good choice of condition can reduce the computations very much. Fock has consistently made use of the so-called harmonic coordinates defined by the four conditions

$$\frac{\partial}{\partial x^\nu}[(-g)^{1/2}g_{\mu\nu}] = 0, \quad g = \det g_{\mu\nu}.$$

These conditions appear to be applicable in virtually all spaces so far considered and define the system of coordinates apart from what is effectively a Lorentz transformation. The indeterminacy of the coordinate frame is therefore just the same as in special relativity, and the interpretation of the work is thus greatly

† This difficulty almost invalidates the original geodesic postulate. For if a particle is a singularity in the field, ds does not exist there, and so it is pointless to speak of a geodesic. For distributed matter ($T^{\mu\nu}$) this difficulty does not arise.

simplified. Using these coordinates, Fock, Petrova and others of his school have made very considerable advances in the theory of the equations of motion and of gravitational waves.

5.3. *Gravitational Waves*

The transition from electrostatics to the theory of varying fields leads to the appearance of electromagnetic waves in the theory. It might therefore be expected that, as static gravitational fields correspond so closely to Coulomb fields in electrostatics, similarly in gravitational theory the consideration of varying fields should lead to gravitational waves. However, the resulting problems turn out to be highly involved, for various reasons.

In the first instance, owing to the weakness of gravitational coupling (compared with electromagnetic coupling) appreciable gravitational fields are due only to large masses which are effectively beyond experimental control and never move in nature with such accelerations as could be expected to produce measurable gravitational waves. In fact, on any order of magnitude argument, gravitational waves are many orders of magnitude below that needed to produce any observable effects.

Secondly, the non-linearity of the field equations raises great mathematical obstacles. These are substantially aggravated by the unlimited freedom to change coordinates which makes it hard to get hold of any significant expressions.

Thirdly, the physical dissimilarity with electromagnetic waves is far greater than appears at first sight. The generation of electromagnetic waves is restricted only by the law of conservation of charge, and there is no need to specify the electrically neutral part of the transmitter since it is irrelevant. In gravitational theory the operation of the transmitter is restricted by numerous conservation laws, and the entire transmitter has to be specified in detail since every part of it contributes to the wave. There are also some difficulties about the receiver which are due to the fact that only the curvature tensor, through the equation of geodesic variation (II.3), is capable of generating heat in a local receiver and thus the *second* derivatives (and not, as in electromagnetism, the first derivatives) of the potentials are physical variables.

Although gravitational waves are so weak, the subject is well worth studying. For not only is it intrinsically desirable to gain information about all the consequences of an otherwise successful theory, but the question of the existence of true gravitational waves representing independent degrees of freedom is vital to the problem of quantization. If gravitation is merely an aura surrounding matter, i.e. a Coulomb type of inter-particle force, then quantization is not needed, but if there are gravitational waves representing independent degrees of freedom, gravitation, like all other fields, requires quantization, or otherwise it conflicts with the uncertainty principle.

The wave problem divides naturally into two parts: propagation of information and propagation of energy. The first has been successfully solved by Lichnerowicz, who was able to show that the empty space field equations (II.4) form a normal hyperbolic system of equations, and to formulate the Cauchy problem appropriate to it. This great step forward has not only clarified the content of field equation (II.4), but has helped Lichnerowicz to discuss the admissible

forms of $T^{\mu\nu}$ in equation (II.5) and so to shed some light on the important problem posed by the absence of negative mass in nature. Pirani has been able to specify the structure of the curvature tensor in a wave zone (as opposed to Coulomb-like regions) and has thus given a clear definition of wave character.

The propagation of energy is far less well understood. There are reasons to believe that the obvious approach by means of a linearized theory and successive approximations may not always be convergent. Exact solutions are known for plane progressive waves and cylindrical waves, but it is quite likely that true three-dimensional waves (such as would be emitted by a finite system) have rather different properties. Though some progress has been made with constructing approximate three-dimensional wave solutions, these are not easily interpreted. The whole problem of whether there is a one-parameter family of solutions from 'retarded' to 'advanced' has hardly been touched. Though it is by now plausible that there are gravitational waves in general relativity it has not yet been proved beyond doubt that a finite system of matter-in-motion radiates energy.

5.4. *The Hamiltonian Formulation and Quantization*

As has already been said, if gravitational waves exist, the gravitational field needs to be quantized to avoid a logical contradiction. Moreover, it is widely held that such a quantization would throw light on the intractable problem of elementary particles, since gravitation is one of their inalienable properties. The usual approach to a quantum theory is by way of a Hamiltonian. In spite of great efforts the Hamiltonian formulation of general relativity is still not quite clear, mainly because the freedom to change coordinates introduces great complications. However, considerable progress towards such a formulation has been made.

The difficulty in finding an acceptable Hamiltonian formulation has retarded advances towards a quantum theory of gravitation but several attempts have come fairly close to success.

5.5. *Cosmology*

Very early in the development of general relativity Einstein realized that the theory was linked with the problems of the structure of the universe, in the first instance through the question of boundary conditions. Led by certain ideas since abandoned about cosmology, he added the term $\lambda g_{\mu\nu}$ (where λ is the so-called cosmological constant) to the left-hand side of (II.5) and correspondingly to the right-hand side of (II.4). These augmented field equations, together with the postulate of the large-scale spatial uniformity of the universe (the so-called cosmological principle) forms the basis of relativistic cosmology. As Friedmann showed, under these conditions there exist a number of model universes. Almost all of them show at least at some stage of their evolution the characteristic expansion associated with observations of distant regions. While in this way the smoothed-out models are readily represented by the theory, the problem of the physical content of a model with the evolution of its irregularities and the nuclear problems posed by a possible early highly condensed state are very difficult indeed. These problems have been considered more for some models than for others, and in particular the work of Lemaitre, and also of Gamow and his collaborators, has gone some way towards examining these questions. Much remains to be done

before it can be said to which if any of these models our actual universe corresponds.

As far as large-scale geometry is concerned, some of these models are of spatially finite extent (positive curvature), including the most important model due to Lemaitre, others are infinite in extent.

Although the work of McCrea, Milne, Layzer, Heckmann and Schücking has demonstrated the existence of exact Newtonian analogues to the smoothed-out relativistic models, it yet remains true to say that relativity is of vital significance to cosmology. For only relativity is able to treat in a logically complete way an unbounded distribution of matter, without the need to introduce special limiting processes or other devices. In so far as the spatial homogeneity of the universe can be said to be established, the unique ability of general relativity to deal directly with such systems may be regarded as a further proof of the superiority of relativity to Newtonian theory, additional to the three crucial tests.

Relativity is not necessarily, however, the best starting-point for cosmology, and the steady-state theory of cosmology in particular draws on other ideas, and arrives at conclusions markedly at variance with those of any of the models of relativistic cosmology. Only observation can establish its claim to be a correct representation of the smoothed-out universe, but the theory cannot be regarded as complete until it is in some way linked to general relativity. For only then will it give a description not merely of the universe as a whole, but also of its large-scale features. The establishment of such a link is not likely to present insuperable difficulties, as the pioneer attempts of Hoyle and McCrea show. In the event of substantial observational support for the steady-state theory the nature of the link may throw further light on general relativity as such.

Cosmological problems are also involved in the vexed question of Mach's principle. This regards as causal the coincidence between the local inertial frame and that within which the universe appears to be isotropic and non-rotating. The cosmological models account for this fact implicitly in the smoothed-out case, but it seems true that general relativity without further assumptions cannot wholly account for Mach's principle. It is likely, however, that the cosmological principle in a suitable form (Weyl's postulate?) is a sufficient assumption to enable Mach's principle to be incorporated into the theory. The work of Sciama has greatly furthered understanding of the questions involved, but yet clearer formulations of these problems are required.

In a very fundamental way Mach's principle may be regarded as the only *explanation* of the principle of equivalence. For it effectively means that inertia is due to the gravitational action of the distant masses of the universe. Thus the equal response of matter to inertia and to gravitation becomes clear. In this way it should also be possible to account for the value of the constant of gravitation in terms of the structure of the universe.

5.6. *Electromagnetism*

So far general relativity has been considered entirely as a theory of inertia and gravitation. There is no difficulty whatever in writing Maxwell's equations in general relativistic form and thus incorporating them into the theory. Rather surprisingly, however, the wave equation for the electromagnetic field so derived contains the curvature tensor. While this is irrelevant for ray optics (since the

second derivatives are not involved) it does affect wave optics and again shows that the principle of equivalence does not hold for sufficiently refined experiments.

In this formulation electromagnetic fields are *in* space while the gravitational field is *of* space. This has seemed unsatisfactory to a number of physicists who have felt that, if going from Euclidean to Riemannian space abolished gravitation as a separate force, a further generalization of the geometry might abolish electromagnetism as a force and thus entirely geometrize the whole of classical physics. In spite of the work done by Weyl, Eddington, Jordan, Einstein and Schrödinger, these attempts to produce a so-called unified field theory based on a super-Riemannian geometry have not led to satisfactory conclusions. No new experiments have been predicted, no further insight into classical physics has been gained, and the hopes of making a contribution to the elementary particle problem have also been disappointed.

An entirely different approach is due to Wheeler. In the first instance he made use of the discovery of Rainich that the $T^{\mu\nu}$ due to a pure electromagnetic field was so specific that one could infer the electromagnetic field from a knowledge of the metric. In this sense he regards general relativity as an 'already unified' theory. In this matter-free space one can now imagine electromagnetic (or even gravitational) waves captured by the gravitational field of their own energy. These objects are called *geons* and from outside appear exactly like matter. Next Wheeler and Misner noticed that Riemannian geometry was compatible with a large range of topologies, and that the introduction of suitable linkages ('wormholes') leads to the appearance of places in space externally indistinguishable from electric charges. Thus the Wheeler–Misner theory of geometrodynamics shows that a Riemannian geometry of suitable topology satisfying everywhere the Rainich empty-space field equations represents a system behaving as though it contained masses and charges. This beautiful theory thus deals satisfactorily with the whole of classical physics without ever introducing mass or charge. It has given new insight into the importance of topology to physics. It has not, however, made as yet any contribution to the elementary particle problem.

Acknowledgments

In conclusion, I wish to express my indebtedness to P. Bergmann, V. Fock, T. Gold, C. W. Kilmister, D. Layzer, W. H. McCrea, F. A. E. Pirani, A. Trautman and J. A. Wheeler. My ideas have been shaped in discussion with these physicists and through reading their papers, but the responsibility for the views here presented is entirely mine. My thanks are also due to the U.S. Air Force whose support has enabled several of these discussions to be held.

References

There are numerous textbooks of special relativity. Particular mention should be made of:

McCrea, W. H., *Relativity Physics* (London: Methuen 1955).

Stephenson, G., and Kilmister, C. W., *Special Relativity for Physicists* (London: Longmans 1958).

Synge, J. L., *Relativity* (Amsterdam: North Holland 1958).

The clock paradox has recently been discussed by H. BONDI (*Discovery*, 1957, December, 505), where references to earlier papers are given.

General relativity is described in numerous textbooks. Special mention should be made of an article by B. HOFFMAN (*Rev. Mod. Phys.*, 1932, **4**, 173).

The conference on the role of gravitation in physics (University of North Carolina, Chapel Hill) in January 1957 discussed most modern problems in general relativity. The proceedings of the conference (WADC *Technical Report*, 57–216) and the associated papers (*Rev. Mod. Phys.*, 1957, **29**, 255–546) contain so much information that reference need only be made here to later papers. The most important of those known to me are :

MISNER, C., and WHEELER, J. A., 1957, *Annals of Physics*, **3**.
MARDER, L., *Proc. Roy. Soc.* A, 1958, **244**, 524.
BERGMANN, P. G., WADC, TN, 58–173.
TRAUTMANN, A., *Lectures on General Relativity*, WADC, TN, 58–251.
PIRANI, F. A. E., *Phys. Rev.*, 1957, **105**, 1089.
BONDI, H., *Nature, Lond.*, 1957, **179**, 1072.
DIRAC, P. A. M., *Proc. Roy. Soc.* A, 1951, **246**, 326, 333.

DOUBLE BETA DECAY

By H. PRIMAKOFF and S. P. ROSEN
Department of Physics, Washington University, St. Louis, Missouri, U.S.A.

CONTENTS

	PAGE
§ 1. Introduction	121
§ 2. Discussion of experimental evidence	127
§ 3. Theory of $\beta\beta$ decay	136
3.1. Nucleon–lepton interaction	136
3.2. Single $\beta^{\mp}$ decay	140
3.3. General formulation of the $\beta^{\mp}\beta^{\mp}$ decay transition probability amplitude	143
3.4. No-neutrino $\beta^{\mp}\beta^{\mp}$ decay	143
3.5. Two-neutrino $\beta^{\mp}\beta^{\mp}$ decay	152
3.6. $\beta\beta$ decays with orbital negatron capture	160
3.7. Relation of $\beta\beta$ decay to inverse β decay	160
§ 4. Comparison of experimental and theoretical $\beta\beta$ half-lives	162
§ 5. Conclusion	163
Acknowledgments	164
References	164

Abstract. The phenomenon of double beta decay is discussed rather fully both from the experimental and from the theoretical point of view, the relation between double beta decay and the possible kinds of inverse beta decay being also briefly treated. A summary of the available experimental data is given, all methods of detection of the various double beta decay processes being considered. The theory of double beta decay, both in the no-neutrino and in the two-neutrino cases, is worked out, *ab initio*, on the basis of a nucleon–lepton interaction without conservation of parity. It is particularly emphasized that, with a 'two-component neutrino' type nucleon–lepton interaction without conservation of parity, absence of no-neutrino $\beta\beta$ decay does not by itself uniquely imply that neutrino and anti-neutrino are distinguishable and that the total lepton charge is conserved.

Comparison of the experimental limits on the double beta decay half-lives with the corresponding theoretical values indicates that two neutrinos are emitted together with two electrons in each double beta decay process—this conclusion is, however, not yet certain and must be confirmed by further experimental work. A treatment is given of the implications of the actual occurrence of two-neutrino rather than no-neutrino double beta decay for the problems of neutrino–anti-neutrino identity and conservation of total lepton charge. It is concluded, on the basis of (i) the available double beta decay, inverse beta decay *and* muon decay experimental data and (ii) the provisional assumption of the universal applicability of 'two-component neutrino' type coupling, that a verdict may be tentatively reached in favour of a 'Dirac' neutrino, operationally distinguishable from a 'Dirac' anti-neutrino, and with conservation of total lepton charge valid in all neutrino interactions.

§ 1. Introduction

In the process of double beta decay, the (very long-lived) nucleus (A, Z) disintegrates spontaneously to the isobaric nucleus $(A, Z \pm 2)$ with the simultaneous emission of two electrons—negatrons or positrons—accompanied by either (*a*) *two* anti-neutrinos or neutrinos, or (*b*) *no* anti-neutrinos or neutrinos. Alternative to the emission of two positrons, emission of one positron and absorption of one orbital negatron, or absorption of two orbital negatrons, is always possible energetically. The existence of double beta decay is expected on general theoretical

grounds as a second-order effect engendered by the presence of the nucleon–lepton interaction which gives rise, in first order, to the usual single beta decay. It is obvious that any double beta decay *without* neutrino emission violates the principle of conservation of total lepton 'charge'† (Zeldovich 1953, Konopinski and Mahmoud 1953, Marx 1953, 1954, Lee 1957, Pauli 1957, Enz 1957, Kahana and Pursey 1957, Konuma 1958, Lüders 1958, and below).

The theoretically predicted probabilities for any of the $\beta\beta$ transitions are extremely small—experimentally, no certain observation of any $\beta\beta$ decay process has yet been made. None the less, the $\beta\beta$ decay phenomenon is of considerable basic interest because of its above-mentioned connection with the conservation of total lepton charge and because of the information it may yield regarding the nature of the neutrino and the type of nucleon–lepton interaction. In particular, the study of $\beta\beta$ decay may shed some light on the question whether the (electrically neutral) neutrino ν is in every way *identical* with the (electrically neutral) anti-neutrino $\bar{\nu}$ (as the neutral pion is in every way identical with the neutral anti-pion) or whether ν is in some respect *different* from $\bar{\nu}$ (as the neutral kayon is different from the neutral anti-kayon by virtue of their opposite values of strangeness, or the neutron is different from the anti-neutron by virtue of their opposite baryon 'charges' and opposite magnetic moments). The study of $\beta\beta$ decay may also clarify certain characteristics of the parts both with and without conservation of parity of the nucleon–lepton interaction Hamiltonian.

Now it is clear, in connection with the $\nu, \bar{\nu}$ identity problem, that if every elementary particle reaction $A+B+\ldots\rightarrow C+D+\ldots+\nu$ with transition probability amplitude $\mathscr{T}(A+B+\ldots\rightarrow C+D+\ldots+\nu)$ coexists with the corresponding elementary particle reaction $A+B+\ldots\rightarrow C+D+\ldots+\bar{\nu}$ with transition probability amplitude $\mathscr{T}(A+B+\ldots\rightarrow C+D+\ldots+\bar{\nu})$ equal exactly to $\mathscr{T}(A+B+\ldots\rightarrow C+D+\ldots+\nu)$, then ν and $\bar{\nu}$ are operationally indistinguishable and so must be considered identical, i.e. $\nu \equiv \bar{\nu}$, the 'Majorana' neutrino.‡ If, on the other hand,

$$\mathscr{T}(A+B+\ldots\rightarrow C+D+\ldots+\bar{\nu}) \neq \mathscr{T}(A+B+\ldots\rightarrow C+D+\ldots+\nu)$$

for one or more A, B, ..., C, D, ..., then ν and $\bar{\nu}$ are operationally distinguishable and so must be considered different, i.e. ν, the 'Dirac' neutrino $\not\equiv \bar{\nu}$, the 'Dirac' anti-neutrino. It is to be noted that when $\nu \equiv \bar{\nu}$, the $\nu, \bar{\nu}$ lepton charges, $L_\nu, L_{\bar{\nu}} = -L_\nu$, must also satisfy $L_\nu = L_{\bar{\nu}}$ so that $L_\nu, L_{\bar{\nu}}$ vanish; the principle of conservation of total lepton charge is then violated in any process where, for example, an odd number of electrons $e^{\mp}(L_{e^-} = -L_{e^+} = 1)$ accompanied by one or more neutrinos are created and no leptons are destroyed, e.g. in single β decay,

$$\left.\begin{aligned} n &\rightarrow p + e^- + \nu, \\ p &\rightarrow n + e^+ + \nu. \end{aligned}\right\} \qquad \ldots\ldots (1a)$$

† The total lepton 'charge' of a system is defined as the difference between the total number of leptons and the total number of anti-leptons present in the system. The definition of the total baryon 'charge' of a system is entirely analogous. Any lepton which is identical with the corresponding anti-lepton is assigned zero lepton 'charge' (see below).

‡ To be completely explicit, if $\nu \equiv \bar{\nu}$ one must have for A, B, ..., C, D, ... :

$$\mathscr{T}(A+B+\ldots\rightarrow C+D+\ldots+\nu) = \mathscr{T}(A+B+\ldots\rightarrow C+D+\ldots+\bar{\nu})$$

for all possible $p_A, p_B, \ldots, p_C, p_D, \ldots, p_\nu, p_{\bar{\nu}} = p_\nu$; $S_A, S_B, \ldots, S_C, S_D, \ldots, S_\nu, S_{\bar{\nu}} = S_\nu$. Here $p_A, \ldots, S_A, \ldots$ are the momenta and spin projections of the indicated particles.

On the other hand, when $\nu \not\equiv \bar{\nu}$ the condition $L_\nu = L_{\bar{\nu}}$ is absent; one has instead only the condition $L_\nu = -L_{\bar{\nu}} = 1$, and the conservation principle in question is consistent with a nucleon–lepton interaction leading, for example, to the single β decay reactions

$$\left.\begin{aligned} &n \to p + e^- + \bar{\nu}, \\ &p \to n + e^+ + \nu. \end{aligned}\right\} \qquad \ldots\ldots (1b)$$

In the $\nu \equiv \bar{\nu}$ or 'Majorana' case, the neutrino magnetic moment is necessarily zero, since for a non-zero magnetic moment the relation between the particle (ν) and anti-particle ($\bar{\nu}$) amplitudes for magnetic moment–photon interaction, $\mathcal{T}(\nu \to \nu' + \gamma) = -\mathcal{T}(\bar{\nu} \to \bar{\nu}' + \gamma)$, contradicts the assumed $\nu, \bar{\nu}$ identity which implies $\mathcal{T}(\nu \to \nu' + \gamma) = \mathcal{T}(\bar{\nu} \to \bar{\nu}' + \gamma)$. Contrariwise, in the $\nu \not\equiv \bar{\nu}$ or 'Dirac' case, ***non-vanishing*** ν and $\bar{\nu}$ magnetic moments, equal in magnitude but opposite in sign, are possible for certain types of nucleon–lepton interaction since here the condition $\mathcal{T}(\nu \to \nu' + \gamma) = \mathcal{T}(\bar{\nu} \to \bar{\nu}' + \gamma)$ need not hold. In fact, theoretical calculations on the basis of a β decay theory with $\nu \not\equiv \bar{\nu}$ and a suitable total lepton charge conserving nucleon–lepton interaction as in equation (1*b*) lead to estimates for the $\nu, \bar{\nu}$ magnetic moments of as much as $\mp 10^{-10 \pm 2}$ electron Bohr magnetons† (Primakoff 1953, Houtermans and Thirring 1954). Experimentally, the magnetic moment of the 'neutrino', ν^-, emitted in the β^- decay of neutron-rich nuclei ($\nu^- = \bar{\nu} \equiv \nu$, equation (1*a*), or $\nu^- = \bar{\nu} \not\equiv \nu$, equation (1*b*)) is less than 10^{-9} electron Bohr magnetons (Cowan and Reines 1957) which is obviously not adequate to differentiate between the Dirac and Majorana cases. However, even if future experiments were to set an upper limit on the neutrino magnetic moment of, say, 10^{-14} electron Bohr magnetons, one could still not unequivocally conclude that $\nu^- = \bar{\nu} \equiv \nu$. Thus, for example, assumption of $\nu^- = \bar{\nu} \not\equiv \nu$ and of total lepton charge conservation, as in equation (1*b*), nevertheless leads to a *zero* neutrino magnetic moment if the (four component) neutrino quantized field amplitude ψ_ν always appears in the nucleon–lepton interaction and in all other neutrino interactions in the 'utmost parity non-conserving' manner appropriate to the 'two-component neutrino', i.e. always appears in the combination $(1+\gamma_5)\psi_\nu$, or in the combination $(1-\gamma_5)\psi_\nu$ (Salam 1957, §3.1 below).

Another choice between the rival possibilities: $\nu \not\equiv \bar{\nu}$ and $\nu \equiv \bar{\nu}$ is possible in principle as a result of the examination of the $e^\pm$ momentum spectrum in muon decay: $\mu^\pm \to e^\pm +$ two 'neutrinos'; the shape of this momentum spectrum is defined by a single numerical parameter, Michel's ρ, with $0 \leqslant \rho \leqslant 1$ (Michel 1950, Bouchiat and Michel 1957, Kinoshita and Sirlin 1957 a, b, Larsen *et al.* 1957).

† The types of nucleon–lepton interaction which are used in the estimates are also such as to maximize (for given magnitudes of the β decay coupling constants) the size of the $\nu, \bar{\nu}$ magnetic moments. The calculation proceeds according to the scheme

$$\begin{array}{ccccccc} \nu & \longrightarrow & p + \bar{n} + e^- & \longrightarrow & p' + \bar{n}' + e^{-\prime} & \longrightarrow & \nu' \\ & \text{virtual dissociation} & & \text{interaction with external magnetic field} & & \text{virtual recombination} & \\ \bar{\nu} & \longrightarrow & \bar{p} + n + e^+ & \longrightarrow & \bar{p}' + n' + e^{+\prime} & \longrightarrow & \bar{\nu}' \end{array}$$

and one has: neutrino magnetic moment $\simeq$ {fraction of time spent by neutrino as a virtual [negatron + anti-neutron + proton]} $\times$ {magnetic moment of negatron} $\simeq -$ {[dimensionless beta decay coupling constant]^2 $\times$ [ratio of nucleon mass to electron mass]^2} $\times$ {electron Bohr magneton} $\simeq -10^{-9}$ $\{e\hbar/2mc\}$.

If now $\frac{3}{4} < \rho \leqslant 1$, the two 'neutrinos' are *necessarily* ν and $\bar{\nu}$ with $\nu \not\equiv \bar{\nu}$, whereas, if $0 \leqslant \rho \leqslant \frac{3}{4}$, one may have as the two 'neutrinos' either $\nu, \bar{\nu}$ or ν, ν or $\bar{\nu}, \bar{\nu}$ all with $\nu \not\equiv \bar{\nu}$, or ν, ν with $\nu \equiv \bar{\nu}$. Experimentally ρ seems to be somewhat less than $\frac{3}{4}$ (Sargent *et al.* 1955, Crowe 1957, Dudziak and Sagane 1957, Rosenson 1958) so that again, unless further assumptions are made, no decision can be reached between the Dirac and the Majorana possibilities. On the other hand, in a theory with 'two-component neutrino' type coupling, the $e^{\pm}$ momentum spectrum from $\mu^{\pm} \to e^{\pm} + \nu + \bar{\nu}$, with $\nu \not\equiv \bar{\nu}$, is characterized by $\rho = \frac{3}{4}$, while the $e^{\pm}$ momentum spectrum from $\mu^{\pm} \to e^{\pm} + \nu + \nu$ (or $\mu^{\pm} \to e^{\pm} + \bar{\nu} + \bar{\nu}$), with $\nu \not\equiv \bar{\nu}$ or $\nu \equiv \bar{\nu}$, is characterized by $\rho = 0$ (Lee and Yang 1957, Salam 1957, Landau 1957, Bouchiat and Michel 1957, Kinoshita and Sirlin 1957a, b, Larsen *et al.* 1957). Thus, if on the basis of further theoretical and experimental work it becomes quite certain that the 'two-component neutrino' type coupling is indeed applicable to muon decay, already available experimental data ($\rho \simeq \frac{3}{4}$) will indicate that $\nu \not\equiv \bar{\nu}$ and so will exclude the $\nu \equiv \bar{\nu}$ Majorana possibility.

We shall now describe in a schematic way the relation between the character of the $\beta\beta$ decay and the question of the identity of ν and $\bar{\nu}$. Suppose, first, that $\nu \equiv \bar{\nu}$; then the possible single β decay reactions are :

$$n \to p + e^- + \nu, \qquad \ldots\ldots(2a)$$

$$\nu + n \to p + e^-, \qquad \ldots\ldots(2b)$$

$$p \to n + e^+ + \nu, \qquad \ldots\ldots(2c)$$

$$\nu + p \to n + e^+, \qquad \ldots\ldots(2d)$$

so that the $\beta\beta$ decay can in general proceed *without* neutrino emission (Racah 1937, Furry 1939), for example,

$$n_1 + n_2 \xrightarrow[\text{via eq. (2a)}]{} n_1 + p_2 + e_2^- + \nu \xrightarrow[\text{via eq. (2b)}]{} p_1 + e_1^- + p_2 + e_2^-, \qquad \ldots\ldots(3)$$

the virtual neutrino emitted by one neutron being absorbed by the other neutron. $\beta\beta$ decay without neutrino emission is, in general, also possible when $\nu \not\equiv \bar{\nu}$ provided that the nucleon–lepton interaction violates the principle of conservation of total lepton charge, i.e. provided that the nucleon–lepton interaction is such that the single β decay reactions

$$n \to p + e^- + \bar{\nu}, \qquad \ldots\ldots(4a)$$

$$\nu + n \to p + e^-, \qquad \ldots\ldots(4b)$$

$$p \to n + e^+ + \nu, \qquad \ldots\ldots(4c)$$

$$\bar{\nu} + p \to n + e^+ \qquad \ldots\ldots(4d)$$

and

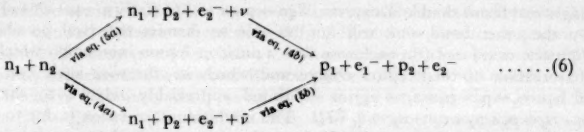
$$n \to p + e^- + \nu, \qquad \ldots\ldots(5a)$$

$$\bar{\nu} + n \to p + e^-, \qquad \ldots\ldots(5b)$$

$$p \to n + e^+ + \bar{\nu}, \qquad \ldots\ldots(5c)$$

$$\nu + p \to n + e^+ \qquad \ldots\ldots(5d)$$

both proceed with *non-vanishing* and more or less comparable transition probabilities.† For in this case, one will again in general have a non-vanishing transition probability for



.....(6)

so that $\beta\beta$ decay without neutrino emission can occur (Touschek 1948, Tiomno 1950).

On the other hand, *only* $\beta\beta$ decay with emission of *two* neutrinos can proceed if $\nu \not\equiv \bar{\nu}$ and if the nucleon–lepton interaction conserves the total lepton charge so that *only* the single β decay reactions in equations (4) proceed with non-vanishing transition probabilities.

Under these circumstances, the $\beta\beta$ decay process involves (Mayer 1935), for example,

$$n_1 + n_2 \xrightarrow[\text{via eq. (4a)}]{} n_1 + p_2 + e_2^- + \bar{\nu}_2 \xrightarrow[\text{via eq. (4a)}]{} p_1 + e_1^- + \bar{\nu}_1 + p_2 + e_2^- + \bar{\nu}_2, \qquad \text{.....(7)}$$

it being here impossible for the anti-neutrino emitted by one neutron to be absorbed by the other neutron. It is also clear that the two-neutrino $\beta\beta$ decay of equation (7) may occur *alternatively* to the no-neutrino $\beta\beta$ decay of equation (6) or equation (3) whenever this no-neutrino $\beta\beta$ decay may itself occur, as for example when $\nu \equiv \bar{\nu}$. It should, however, be noted that when two-neutrino and no-neutrino $\beta\beta$ decay are both possible and are both associated with transition matrix elements of comparable size, the latter process has, in general, a far greater probability than the former—this follows since the available phase space volume is much larger for the virtual neutrino emitted and re-absorbed in the no-neutrino $\beta\beta$ process than for the real neutrinos emitted in the two-neutrino $\beta\beta$ process. Thus, if no-neutrino $\beta\beta$ decay can occur at all, its rate is in general expected to be considerably greater than the rate of two-neutrino $\beta\beta$ decay, when the latter is the only possibility. In addition, in no-neutrino $\beta\beta$ decay the sum of the kinetic energies of the two emitted electrons $\epsilon_1 + \epsilon_2$ is the same in each and every disintegration process; on the other hand, in two-neutrino $\beta\beta$ decay, $\epsilon_1 + \epsilon_2$ is different for different disintegrations, the emitted neutrinos carrying off, on the average, roughly as much energy as the electrons. Experimentally, attempts to distinguish no-neutrino $\beta\beta$ decay from two-neutrino $\beta\beta$ decay have been based on this *expected difference in the* $\epsilon_1 + \epsilon_2$ *distributions*, as well as on the above-mentioned *expected difference in the* $\beta\beta$ *lifetimes* for a given $(A, Z) \rightarrow (A, Z \pm 2)$ available energy (see §2, below). Once unequivocal $\beta\beta$ decay measurements are obtained, the position will be as follows:

† It should be emphasized that if the transition probability amplitudes for the various reactions in equations (4) and the corresponding reactions in equations (5) are exactly equal (i.e. if $\mathcal{T}(n \rightarrow p + e^- + \bar{\nu}) = \mathcal{T}(n \rightarrow p + e^- + \nu)$; $\mathcal{T}(\nu + n \rightarrow p + e^-) = \mathcal{T}(\bar{\nu} + n \rightarrow p + e^-)$; etc.), then the ν and $\bar{\nu}$ transition probability amplitudes for at least one other elementary particle reaction must be different if ν is to be $\not\equiv \bar{\nu}$ (e.g. one must have $\mathcal{T}(\pi^+ \rightarrow \mu^+ + \nu) \neq 0$, $\mathcal{T}(\pi^+ \rightarrow \mu^+ + \bar{\nu}) = 0$; $\mathcal{T}(\pi^- \rightarrow \mu^- + \bar{\nu}) \neq 0$, $\mathcal{T}(\pi^- \rightarrow \mu^- + \nu) = 0$).

I. If observations on $\beta\beta$ decay establish its *two-neutrino* character, it will be simplest to conclude that: (*a*) the neutrino and anti-neutrino are distinguishable, $\nu \not\equiv \bar{\nu}$, and the nucleon–lepton interaction conserves the total lepton charge in single and hence double decay, i.e. $\mathscr{T}(n \to p + e^- + \bar{\nu}) \neq 0$, $\mathscr{T}(n \to p + e^- + \nu) = 0$, etc. On the other hand, one will not be able to dismiss the rival possibility that (*b*) $\nu \not\equiv \bar{\nu}$, or (*c*) $\nu \equiv \bar{\nu}$, in each case with a nucleon–lepton interaction which violates conservation of total lepton charge and which is otherwise such that the rate of $n_1 + n_2 \to p_1 + p_2 + e_1^- + e_2^-$ is inhibited appreciably relative to the rate of $n_1 + n_2 \to p_1 + p_2 + e_1^- + e_2^- + \bar{\nu}_1 + \bar{\nu}_2$. The inhibition in question is due to cancellation in the relevant no-neutrino $\beta\beta$ decay transition matrix element; it is important to mention that the 'two-component neutrino' type coupling results in complete cancellation both in the $\nu \not\equiv \bar{\nu}$ and in the $\nu \equiv \bar{\nu}$ case, and hence gives only two-neutrino $\beta\beta$ decay (§ 3.3, below). In general, however, it is clear that if this cancellation is to inhibit successfully the no-neutrino $\beta\beta$ decay compared with the two-neutrino $\beta\beta$ decay, the appropriate inhibition factor must be big enough to overcome the effect of the above-mentioned large numerical value for the ratio of virtual to real neutrino phase space volumes, which always favours the no-neutrino $\beta\beta$ process.

II. If observations on $\beta\beta$ decay establish its *no-neutrino* character, it will be necessary to conclude that the principle of conservation of total lepton charge is violated in the nucleon–lepton interaction and that either: (*a*) $\nu \not\equiv \bar{\nu}$ with both $\mathscr{T}(n \to p + e^- + \bar{\nu})$ and $\mathscr{T}(n \to p + e^- + \nu)$, etc., non-vanishing; or (*b*) $\nu \equiv \bar{\nu}$. It thus follows that a decision on the basis of experiment in favour of either two-neutrino or no-neutrino $\beta\beta$ decay is, *of itself*, not sufficient to choose between the Dirac ($\nu \not\equiv \bar{\nu}$) and the Majorana ($\nu \equiv \bar{\nu}$) possibilities. Nevertheless, such an experimental decision would certainly provide valuable information towards documenting any eventual choice between these two and, at least if $\beta\beta$ decay turned out to be the no-neutrino type, would unequivocally demonstrate that conservation of total lepton charge is violated.

We conclude this introductory section with a brief treatment of the necessarily intimate relation between the character of $\beta\beta$ decay and the type of inverse β decay. With ν^-, ν^+ the 'neutrinos' emitted in the β^-, β^+ decay of neutron-rich, proton-rich nuclei, the possible inverse β decay reactions are:

$$\nu^- + p \to n + e^+, \qquad \ldots\ldots(8)$$

$$\nu^+ + n \to p + e^-, \qquad \ldots\ldots(9)$$

$$\nu^- + n \to p + e^-, \qquad \ldots\ldots(10)$$

$$\nu^+ + p \to n + e^+. \qquad \ldots\ldots(11)$$

Using the notation immediately above, we distinguish between the following cases:

(A): reactions (8), (9) allowed; reactions (10), (11) forbidden. Under these circumstances, no-neutrino $\beta\beta$ decay is not possible as a rule, and two-neutrino $\beta\beta$ decay is, in general, anticipated: I(*a*) with $\{\nu^- = \bar{\nu}\}$, $\{\nu^+ = \nu \not\equiv \bar{\nu}\}$; or I(*b*) with $\{\nu^- = \bar{\nu}$ or $\nu \not\equiv \bar{\nu}\}$, $\{\nu^+ = \nu$ or $\bar{\nu} \not\equiv \nu\}$; or I(*c*) with $\{\nu^- = \bar{\nu} \equiv \nu = \nu^+\}$.

(B): reactions (8), (9) allowed; reactions (10), (11) also allowed. Under these circumstances, no-neutrino $\beta\beta$ decay is, in general, anticipated: II(a) with $\{\nu^- = \tilde{\nu}$ or $\nu \not\equiv \tilde{\nu}\}$, $\{\nu^+ = \nu$ or $\tilde{\nu} \not\equiv \nu\}$; or II($b$) with $\{\nu^- = \tilde{\nu} \equiv \nu = \nu^+\}$.

We therefore see in effect that a one-to-one correspondence exists between the forbidden or allowed character of reactions (10), (11), and the absence or presence of no-neutrino $\beta\beta$ decay (see §3.7, below). The reaction (8) has been definitely established by experiment (Cowan and Reines 1953a, b, 1956) while reaction (10) is under intensive laboratory study using as the target neutrons those bound in $^{37}_{17}\mathrm{Cl}$: $\nu^- + {}^{37}_{17}\mathrm{Cl} \rightarrow {}^{37}_{18}\mathrm{Ar} + \mathrm{e}^-$ (Davis 1955, 1956, 1957 private communication); so far only an upper limit for the rate of reaction (10) has been reported (Davis 1957, private communication). A review of the inverse β decay experiments, as well as a brief discussion of $\beta\beta$ decay, has been given by Ridley (1956) in an excellent article written before the discovery of the non-conservation of parity in the nucleon–lepton interaction. A more recent review of inverse β decay as well as of $\beta\beta$ decay is contained in a monograph on the neutrino by Allen (1958).

§2. Discussion of Experimental Evidence

Considerations of nuclear stability indicate that both the parent (A, Z) and daughter $(A, Z \pm 2)$ nuclei in all $\beta\beta$ decay processes likely to be observed are of the even–even type. Thus the initial state in any such process, being the ground state of an even–even nucleus, is characterized by zero spin and positive parity, 0^+; the final state is normally expected to be the ground state of the daughter nucleus, which is again 0^+—however $\beta\beta$ decay transitions to energetically accessible excited states of the daughter, e.g. to a 2^+ first excited state, may also be anticipated.

The 'intermediate' nucleus $(A, Z \pm 1)$, 'situated between' an observable $\beta\beta$ parent (A, Z) and the corresponding $\beta\beta$ daughter $(A, Z \pm 2)$, is necessarily such that either:

(a) The single β decay $(A, Z) \rightarrow$ any state of $(A, Z \pm 1)$ is not possible energetically; or

(b) The single β decay $(A, Z) \rightarrow (A, Z \pm 1)$ is possible energetically, for instance to the ground state of $(A, Z \pm 1)$, but is so strongly inhibited by an associated large spin change that the $\beta\beta$ decay $(A, Z) \rightarrow (A, Z \pm 2)$ is actually more probable. An example of case (b) occurs in the triplet $^{48}_{20}\mathrm{Ca}$, $^{48}_{21}\mathrm{Sc}$, $^{48}_{22}\mathrm{Ti}$ (see below); here the $\beta^-\beta^-$ decay process

$$0^+ \text{ ground state of } {}^{48}_{20}\mathrm{Ca} \rightarrow 0^+ \text{ ground state of } {}^{48}_{22}\mathrm{Ti}$$

is expected to proceed at a much faster rate than the energetically possible but sixth forbidden single β^- decay

$$0^+ \text{ ground state of } {}^{48}_{20}\mathrm{Ca} \rightarrow 6 \text{ ground state of } {}^{48}_{21}\mathrm{Sc}.$$

The 'energy release' in a $\beta\beta$ decay process is equal to the maximum kinetic energy of either of the emitted electrons $e_1^{\mp}$, $e_2^{\mp}$, provided that the mass of the neutrino is taken as zero and the recoil energy of the daughter nucleus is neglected. By conservation of energy, this 'energy release' (divided by c^2) must then be equal to the atomic mass difference between the $\beta\beta$ parent and daughter in the double negatron case, and the atomic mass difference less four electron masses, i.e. less $2\,\mathrm{Mev}/c^2$, in the double positron case. Further, in a $\beta\beta$ decay process

where one positron is emitted and one orbital (K-shell) negatron with binding energy η_K is captured, the 'energy release', as given by the maximum kinetic energy of the emitted e^+, is the $\beta\beta$ parent–daughter mass difference ($\times c^2$) less $\{1\,\text{MeV}+\eta_K\}$; if this $\beta\beta$ decay is no-neutrino, the e^+ is mono-energetic. Finally, in a $\beta\beta$ process where two orbital (K-shell) negatrons are captured, the energy release, as given by the maximum energy of either of the emitted neutrinos, or, if the decay is of no-neutrino type, by the energy of the emitted (mono-energetic) inner bremsstrahlung photon, is just the parent–daughter atomic mass difference ($\times c^2$) less $2\eta_K$.

Experiments on $\beta\beta$ decay fall into three main categories :

(*a*) Detection of the daughter element $(A, Z\pm 2)$ of the decay, by a study of its abundance within sources of the parent (A, Z).

(*b*) Detection in a cloud chamber, or photographic emulsion of $e_1^{\mp}, e_2^{\mp}$ tracks originating from a common point within the $\beta^{\mp}\beta^{\mp}$ decaying element.

(*c*) Counter detection of $e_1^{\mp}, e_2^{\mp}$ time-coincidences from a $\beta^{\mp}\beta^{\mp}$ decaying element.

It should be emphasized that method (*a*) can yield only the rate, or half-life of the decay, while methods (*b*), (*c*) can yield not only the half-life, but also the $e_1^{\mp}, e_2^{\mp}$ kinetic energy distributions, in particular that for the sum of the kinetic energies of $e_1^{\mp}$ and $e_2^{\mp}$. Up to the present, only method (*c*) has been used in the attempted determination of these distributions.

We begin with a discussion of method (*a*). Here rocks of known constitution and age, and containing the anticipated $\beta\beta$ decay parent elements (A, Z), are analysed by chemical and mass spectrographic methods for the corresponding daughter elements $(A, Z\pm 2)$. Thus in an examination of tellurium ore (70% Bi_2Te of age $1{\cdot}5\times 10^9$ years) containing the parent $^{130}_{52}$Te, Inghram and Reynolds (1949, 1950) found abnormally large isotopic abundance of $^{130}_{54}$Xe in the occluded xenon and consequently estimated a half-life of $1{\cdot}4\times 10^{21}$ years for the $\beta^-\beta^-$ decay $^{130}_{52}\text{Te}\rightarrow{}^{130}_{54}\text{Xe}$. From the measured atomic mass difference between $^{130}_{52}$Te and $^{130}_{54}$Xe, the energy release in this decay is 3·0 MeV (see figure 1), a value which makes the above half-life more consistent with theoretical estimates appropriate to a two-neutrino rather than a no-neutrino $\beta^-\beta^-$ decay. (See §§3, 4 for a discussion of all theoretical estimates for the half-life of $\beta\beta$ decay.) The theoretical estimate for the half-life of the fifth forbidden single β^- decay $^{130}_{52}\text{Te}\rightarrow{}^{130}_{53}\text{I}$, spin $0^+\rightarrow 6^-$, energy release 0·05 MeV, using the semi-empirical formulae due to Feenberg (1955), is of the order of 10^{29} years; if, however, the energy release in the $\beta^-\beta^-$ decay $^{130}_{52}\text{Te}\rightarrow{}^{130}_{54}\text{Xe}$ is not actually 3·0 MeV, but, say, 3·2 MeV, then the energy release in the β^- decay $^{130}_{52}\text{Te}\rightarrow{}^{130}_{53}\text{I}$ is 0·25 MeV, and the estimated half-life is reduced by a factor 10^6 to 10^{23} years. Thus the possibility cannot be dismissed (Kohman 1954) that the excess $^{130}_{54}$Xe observed is produced by the successive single β^- decays :

$$^{130}_{52}\text{Te}\rightarrow{}^{130}_{53}\text{I}+e^-+\bar{\nu},$$

$$^{130}_{53}\text{I}\rightarrow{}^{130}_{54}\text{Xe}+e^-+\bar{\nu}.$$

It should also be mentioned that Inghram and Reynolds found abnormally large isotopic abundances of ^{129}Xe and ^{131}Xe; these, however, are ascribed to the

reactions :

$$^{128,\,130}_{52}\mathrm{Te} + \mathrm{n} \rightarrow {}^{129,\,131}_{52}\mathrm{Te} + \gamma,$$

$$^{129,\,131}_{52}\mathrm{Te} \rightarrow {}^{129,\,131}_{53}\mathrm{I} + \mathrm{e}^- + \bar{\nu},$$

$$^{129,\,131}_{53}\mathrm{I} \rightarrow {}^{129,\,131}_{54}\mathrm{Xe} + \mathrm{e}^- + \bar{\nu},$$

the neutron 'flux' resulting from fission of uranium found in proximity to the tellurium ore.

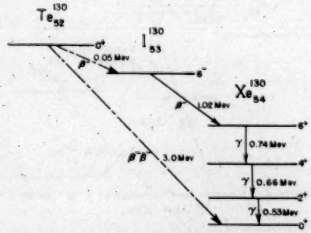


Figure 1. Energy level scheme of $\beta^-\beta^-$ decay : $^{130}_{52}\mathrm{Te} \rightarrow {}^{130}_{54}\mathrm{Xe}$.

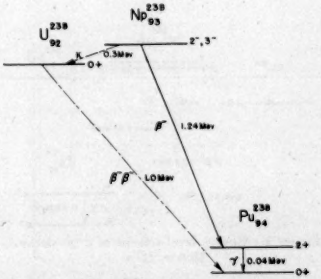


Figure 2. Energy level scheme of $\beta^-\beta^-$ decay : $^{238}_{92}\mathrm{U} \rightarrow {}^{238}_{94}\mathrm{Pu}$.

In a similarly conceived experiment, Levene, Ghiorso and Seaborg (1950) searched for $^{238}_{94}$Pu in samples of $^{238}_{92}$U. After chemical separation of the plutonium fraction from the uranium, any $^{238}_{94}$Pu present was to be identified by counter type detection of its 5·5 Mev α-particle ($^{238}_{94}\mathrm{Pu} \rightarrow {}^{234}_{92}\mathrm{U} + {}^{4}_{2}\mathrm{He}$). No α-particles ascribable to $^{238}_{94}$Pu were found, and, as a result, a lower limit of 6×10^{18} years was set on the half-life of the $\beta^-\beta^-$ decay $^{238}_{92}\mathrm{U} \rightarrow {}^{238}_{94}\mathrm{Pu}$. The energy release is 1·0 Mev (see figure 2), making the above lower limit, once more, consistent with theoretical estimates for the half-life of two-neutrino, rather than with those for no-neutrino, $\beta^-\beta^-$ decay.

Proceeding to a discussion of method (*b*), we describe briefly the experiments of Lawson (1951), of Fireman and Schwarzer (1952), of Fremlin and Walter (1952), and of Winter (1955 a). In all these investigations, no events definitely identifiable as $\beta^\mp\beta^\mp$ decay were observed, and only lower limits could be set on the corresponding half-lives ; again, most of these are longer than reasonable theoretical estimates for no-neutrino $\beta^\mp\beta^\mp$ decay, but are not inconsistent with estimates for the two-neutrino decay.

Lawson (1951) used a cloud chamber to study the $\beta^-\beta^-$ decay $^{124}_{50}\mathrm{Sn} \rightarrow {}^{124}_{52}\mathrm{Te}$ (energy release of 2·2 Mev, see figure 3), with a source of 0·268-g, containing $^{124}_{50}$Sn with an isotopic abundance of 83% (natural abundance of $^{124}_{50}$Sn = 6%). He did not observe any pairs of negatron tracks that could be definitely ascribed to $\beta^-\beta^-$ events originating in the ^{124}Sn source, and set a lower limit on the half-life of the $\beta^-\beta^-$ decay of 10^{16} years.

Fireman and Schwarzer (1952) examined the same decay, comparing a 2 g $_{50}$Sn source of natural isotopic abundance with another of 2 g, enriched to 95% $^{124}_{50}$Sn. A helium-filled cloud chamber in a magnetic field, triggered by internal thin-walled counters in time-coincidence, was employed. Of 4000 photographs, 1500 with the natural, and 2500 with the enriched, source, only three contained a pair of electrons appearing to come from the same point in the source, and with each electron entering one of the counters; all three of these photographs were obtained with the enriched source. On the other hand, in all these cases,

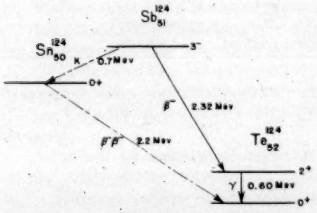


Figure 3. Energy level scheme of $\beta^-\beta^-$ decay : $^{124}_{50}\mathrm{Sn} \rightarrow {}^{124}_{52}\mathrm{Te}$.

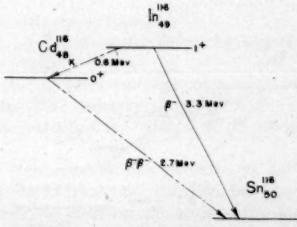


Figure 4. Energy level scheme of $\beta^-\beta^-$ decay : $^{116}_{48}\mathrm{Cd} \rightarrow {}^{116}_{50}\mathrm{Sn}$.

one member of the electron pair, endowed with the 'wrong' curvature due to multiple scattering, could actually have been moving towards, rather than away from, the source. Consequently, Fireman and Schwarzer assigned only a lower limit to the half-life of the $\beta^-\beta^-$ decay, $^{124}_{50}\mathrm{Sn} \rightarrow {}^{124}_{52}\mathrm{Te}$. This lower limit of 10^{17} years is in contradiction to Fireman's earlier observation (method (*c*)) of a half-life of 10^{16} years for this decay (Fireman 1949).

Winter (1955 a) took random pictures in an argon-filled cloud chamber placed in a magnetic field. The sources, both of natural isotopic abundance, were a 30 g foil of $_{48}$Cd and a 42 g foil of $_{42}$Mo, and the processes in table 1 were investigated. Twenty-four pairs of negatron tracks and one pair of positron tracks, in each pair the tracks originating from the same point of the foil, were observed in 12 000 photographs of the Cd source, and thirty-five pairs of negatron tracks in 13 000 photographs of the Mo. All of these pairs of tracks could, however, be ascribed to one or another of the possible background effects, so that only lower limits on the half-lives could be assigned.

Table 1

$\beta^\mp\beta^\mp$ decay	Energy release (Mev) (see figures 4, 5, 6)	Lower limit on half-life (years)
$^{116}_{48}\mathrm{Cd} \rightarrow {}^{116}_{50}\mathrm{Sn}$	2·7	1×10^{17}
$^{100}_{42}\mathrm{Mo} \rightarrow {}^{100}_{44}\mathrm{Ru}$	2·3	3×10^{17}
$^{106}_{48}\mathrm{Cd} \rightarrow {}^{106}_{46}\mathrm{Pd}$	0·9	6×10^{16}

Fremlin and Walter (1952) studied the decay of various nuclei by a photographic emulsion method; gramme amounts of source material, of natural isotopic abundance (except for $_{50}$Sn, which was enriched with $^{124}_{50}$Sn), were embedded

in the recesses of graphite blocks, facing electron sensitive emulsions; exposure times of the order of months were used and counts made of the number of electron tracks crossing the emulsion surface both in regions 'a' exposed to the source, and regions 'b' in contact with the graphite. Essentially negative results were obtained in all cases, with the possible exception of $_{42}$Mo($^{100}_{42}$Mo$\rightarrow^{100}_{44}$Ru); here, in an exposure of 27 days, 88 electron tracks per mm^2 were observed in region 'a' and 54 in region 'b'. Moreover, this electron activity was not correlated with any α-particle activity as might occur if naturally radioactive impurities

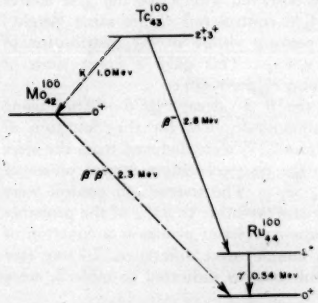


Figure 5. Energy level scheme of $\beta^-\beta^-$ decay: $^{100}_{42}$Mo$\rightarrow^{100}_{44}$Ru.

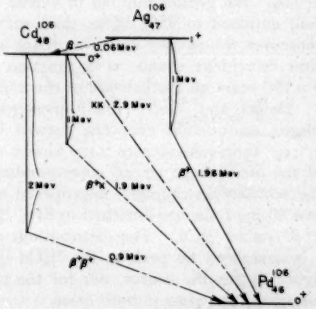


Figure 6. Energy level scheme of $\beta^+\beta^+$ decay: $^{106}_{48}$Cd$\rightarrow^{106}_{46}$Pd.

were present in the source and as was found with sources other than $_{42}$Mo (e.g. $^{124}_{50}$Sn and $_{56}$Ba). The difference between electron track counts, 1·2 electrons per mm^2 per day, in regions 'a' and 'b', for $_{42}$Mo, corresponds, if ascribed to the $\beta^-\beta^-$ decay $^{100}_{42}$Mo$\rightarrow^{100}_{44}$Ru, to a half-life of 2×10^{16} years; this conceivably positive result is, however, to be contrasted with Winter's above-mentioned negative result.

We now treat method (c), employed in the experiments of Kalkstein and Libby (1952), Pearce and Darby (1952), Detoef and Moch (1955), McCarthy (1953, 1955), Awschalom (1956), and Cowan, Harrison, Langer and Reines (1956). No actual observations of $\beta^-\beta^-$ decay events were obtained in any of these investigations apart from McCarthy's positive result on $^{48}_{20}$Ca$\rightarrow^{48}_{22}$Ti, later contradicted by Awschalom, who used similar equipment and operated under more favourable background conditions. Consequently only lower limits, all of them longer than theoretical estimates for no-neutrino $\beta^-\beta^-$ decay, but consistent with those for two-neutrino $\beta^-\beta^-$ decay, may be assigned to the half-lives.

Kalkstein and Libby (1952) searched for e_1^-, e_2^- time-coincidences from the decay $^{124}_{50}$Sn$\rightarrow^{124}_{52}$Te, employing two screen wall counters and a 3 g source enriched to 95% $^{124}_{50}$Sn. A geometrically identical $_{50}$Sn source, of natural isotopic abundance, was used as a control. Within statistical error, the same number of coincidences was observed with both sources and a lower limit of 2×10^{17} years was set on the half-life of the $^{124}_{50}$Sn$\rightarrow^{124}_{52}$Te $\beta^-\beta^-$ decay.

Pearce and Darby (1952) also studied this decay, with counter apparatus to detect e_1^-, e_2^- time-coincidences; in addition they were the first to measure the sum of the kinetic energies $\epsilon_1+\epsilon_2$ of the e_1^-, e_2^- in each time-coincident event observed. The distribution of these events as a function of $\epsilon_1+\epsilon_2$ would peak very sharply at a value equal to the $^{124}_{50}$Sn–$^{124}_{52}$Te atomic mass difference, 2·2 MeV, if they arise from no-neutrino $\beta^-\beta^-$ decays; if, however, the time-coincident events are due to two-neutrino $\beta^-\beta^-$ decay, no marked peaking would be expected.

Scintillation counters (anthracene crystals) were used for the detection of e_1^-, e_2^- and an 18-channel pulse height discriminator for the measurement of $\epsilon_1+\epsilon_2$. No greater number of events were observed with a 200 mg $_{50}$Sn source foil, enriched to 95% $^{124}_{50}$Sn, than with a $^{27}_{13}$Al control foil of the same weight; moreover, in neither case was there any peaking visible in the distribution of time-coincident events as a function of $\epsilon_1+\epsilon_2$. This gave a lower limit of 5×10^{16} years on the half-life of the $\beta^-\beta^-$ decay $^{124}_{50}\mathrm{Sn}\rightarrow{}^{124}_{52}\mathrm{Te}$.

Detoef and Moch (1955) investigated the $\beta^-\beta^-$ decay $^{116}_{48}\mathrm{Cd}\rightarrow{}^{116}_{50}\mathrm{Sn}$, using plastic scintillation counters, viewed by photomultipliers for the detection of e_1^-, e_2^- time-coincidences: the kinetic energies ϵ_1, ϵ_2 were deduced from the sizes of the oscilloscope traces, corresponding to the time-coincident output pulses of the scintillation counters, engendered by e_1^-, e_2^-. The source and control were two 80-mg foils, one enriched to 81% $^{116}_{48}$Cd and the other to 95% of the probably $\beta^+\beta^+$ stable $^{114}_{48}$Cd. The distributions of time-coincident events as a function of $\epsilon_1+\epsilon_2$ showed no peak at the $^{116}_{48}$Cd–$^{116}_{50}$Sn atomic mass difference, 2·7 MeV (see figure 4), for the source, nor for the control, and, as indicated in table 2, were essentially the same in both cases.

Table 2

Sample	$^{116}_{48}$Cd	$^{114}_{48}$Cd
Number of e_1^-, e_2^- time-coincidences in 200 hr with 2 MeV $\leqslant\epsilon_1+\epsilon_2\leqslant$ 3·5 MeV	109	106

The lower limit on the half-life of the $\beta^-\beta^-$ decay $^{116}_{48}\mathrm{Cd}\rightarrow{}^{116}_{50}\mathrm{Sn}$ was set as 6×10^{16} years.

McCarthy (1953) examined the $\beta^-\beta^-$ decay processes $^{124}_{50}\mathrm{Sn}\rightarrow{}^{124}_{52}\mathrm{Te}$ and $^{96}_{40}\mathrm{Zr}\rightarrow{}^{96}_{42}\mathrm{Mo}$ (see also Fulbright 1952, 1955). The energy release in the latter case, calculated from the measured atomic masses of $^{96}_{40}$Zr and $^{96}_{42}$Mo, is 3·3 MeV (see figure 7) and, as indicated in this figure, the single β^- decay $^{96}_{40}\mathrm{Zr}\rightarrow{}^{96}_{41}\mathrm{Nb}$ is also energetically possible unless the $\beta^-\beta^-$ decay energy release is actually less than 3·16 MeV. No definite experimental information on the spin and parity of the ground state of $^{96}_{41}$Nb is available, but an assignment of 6^+ is not inconsistent with nuclear shell model arguments, and with the possible spin values of the excited states of $^{96}_{42}$Mo as indicated by the allowed character of the 0·75 MeV single β^- decay $^{96}_{41}\mathrm{Nb}\rightarrow{}^{96}_{42}\mathrm{Mo}$, and by the absence of crossover γ's in $^{96}_{42}$Mo. Assuming that the $^{96}_{40}\mathrm{Zr}\rightarrow{}^{96}_{41}\mathrm{Nb}$, 0·14 MeV single β^- decay is indeed $0^+\rightarrow6^+$ and so sixth forbidden, its half-life, estimated from Feenberg's (1955) semi-empirical formulae, is of order 10^{26} years. If instead, the energy release of the $\beta^-\beta^-$ decay $^{96}_{40}\mathrm{Zr}\rightarrow{}^{96}_{42}\mathrm{Mo}$ is taken as, say, 3·5 MeV, making the release in the single β^- decay $^{96}_{40}\mathrm{Zr}\rightarrow{}^{96}_{41}\mathrm{Nb}$ equal to 0·34 MeV, then the half-life of the single β^- decay is still of order 10^{24} years.

Thus the estimates of the half-life of this single β^- decay are so much longer than any of those for the $^{96}_{40}Zr \rightarrow {}^{96}_{42}Mo$ $\beta^-\beta^-$ decay (see §§ 3, 4) that the process $^{96}_{40}Zr \rightarrow {}^{96}_{41}Nb$ can be disregarded.

In these experiments of McCarthy, the presumed $\beta^-\beta^-$ source and the control were alternately placed between two *trans*-stilbene scintillation counters, observed by separate photomultipliers for the time-coincident detection of e_1^-, e_2^-. An anti-coincidence system of four large guard counters was used to reduce background. The e_1^-, e_2^- kinetic energies ϵ_1, ϵ_2 were obtained by adding the time-coincident pulses from the two photomultipliers and analysing the resultant pulse

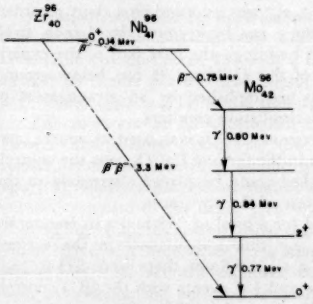


Figure 7. Energy level scheme of $\beta^-\beta^-$ decay: $^{96}_{40}Zr \rightarrow {}^{96}_{42}Mo$.

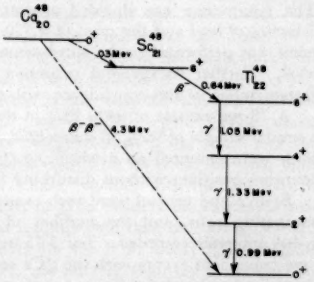


Figure 8. Energy level scheme of $\beta^-\beta^-$ decay: $^{48}_{20}Ca \rightarrow {}^{48}_{22}Ti$.

with a 30-channel pulse-height discriminator. The sources were samples of 95% $^{124}_{50}Sn$ (150 mg) and 90% $^{96}_{40}Zr$ (52 mg), and the controls isotopically natural $_{50}Sn$ (150 mg) and 98% $^{94}_{40}Zr$ (52 mg). $^{94}_{40}Zr$ has a $\beta^-\beta^-$ energy release of 1·4 Mev, and so not only does it have a much longer mean life than $^{96}_{40}Zr$, but it also cannot produce time-coincident events with a kinetic energy sum $\epsilon_1 + \epsilon_2$ in the region of 2·0 to 4·0 Mev.

In runs of some 200 hours, McCarthy found, for all values of $\epsilon_1 + \epsilon_2$, essentially the same number of e_1^-, e_2^- time-coincident events with the sources as with the corresponding controls. Moreover, none of the distributions of these time-coincident events as functions of $\epsilon_1 + \epsilon_2$ showed signs of peaking. Lower limits on the half-lives of the $\beta^-\beta^-$ decays $^{124}_{50}Sn \rightarrow {}^{124}_{52}Te$ and $^{96}_{40}Zr \rightarrow {}^{96}_{42}Mo$ were set at $1{\cdot}5 \times 10^{17}$ and 2×10^{16} years respectively.

After these experiments McCarthy (1955) investigated the $\beta^-\beta^-$ decay $^{48}_{20}Ca \rightarrow {}^{48}_{22}Ti$; this process has an energy release of 4·3 Mev, calculated from the atomic mass difference of $^{48}_{20}Ca$ and $^{48}_{22}Ti$ (figure 8). As seen from this figure, the single β^- decay is energetically possible, with 0·34 Mev release. The spin change $0^+ \rightarrow 6^+$ makes it sixth forbidden, and from Feenberg's (1955) semi-empirical formula, the corresponding half-life is of order 10^{25} years. An energy release of, say, 4·5 Mev, in the process $^{48}_{20}Ca \rightarrow {}^{48}_{22}Ti$, makes that in $^{48}_{20}Ca \rightarrow {}^{48}_{21}Sc$ equal to 0·5 Mev and gives a half-life for the single β^- decay of order 10^{24} years. Both these single

β^- half-life estimates are much longer than any of those for the double $\beta^-\beta^-$ decay $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$ (see §§3, 4) and so, as in the case of $^{96}_{40}\mathrm{Zr}$, the single β^- decay process can be disregarded.

The overall method and geometrical disposition of apparatus in the $^{48}_{20}\mathrm{Ca}$ experiment was similar to that in the earlier experiments on $^{124}_{50}\mathrm{Sn}$ and $^{96}_{40}\mathrm{Zr}$, though some of the equipment was considerably different. To detect e_1^-, e_2^- time-coincidences, two plastic phosphor scintillation counters, each mounted on a photomultiplier and subtending a solid angle of $1{\cdot}8\pi$, were placed on opposite sides of the source. The output pulses of these counters were amplified and added electronically. The resultant pulse height, proportional to the kinetic energy sum $\epsilon_1+\epsilon_2$ of the time-coincident e_1^-, e_2^- was recorded on a chart recorder. The equipment was shielded against cosmic ray background by $\frac{1}{2}$-inch steel, 3 inches of lead and the mass of a 570-foot building—the later part of the experiment was performed in the sub-basement of this building, 35 feet below ground level. Further background reduction was accomplished by an arrangement of Geiger tubes in anti-coincidence with the scintillation counters.

A 78-mg sample of 84% $^{48}_{20}\mathrm{Ca}$ in the form of $CaCO_3$ was used as source, and a similar sample of 98% $\beta\beta$ stable $^{44}_{20}\mathrm{Ca}$, also in the form of $CaCO_3$, was the control. Both were mounted on a wheel, so that they could be placed alternately in the counting position, without disturbing the rest of the apparatus.

Source and control were each examined for a total of 755 hours in frequently alternating runs, and the number of e_1^-, e_2^- time-coincidences in the various $\epsilon_1+\epsilon_2$ intervals recorded. For $3{\cdot}75\ \mathrm{MeV} \leqslant \epsilon_1+\epsilon_2 \leqslant 4{\cdot}5\ \mathrm{MeV}$ there were 595 e_1^-, e_2^- time-coincident events with the $^{48}_{20}\mathrm{Ca}$ source and 455 events with the $^{44}_{20}\mathrm{Ca}$ control. On the other hand, for $2\ \mathrm{MeV} \leqslant \epsilon_1+\epsilon_2 \leqslant 3{\cdot}75\ \mathrm{MeV}$, and for $4{\cdot}5\ \mathrm{MeV} \leqslant \epsilon_1+\epsilon_2 \leqslant 5\ \mathrm{MeV}$, the number of events was essentially the same with source and control. In view of the fact that if no neutrinos are emitted in the decay $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$, the e_1^-, e_2^- time-coincident events distribution would be very sharply peaked at the energy release value of 4·3 MeV, McCarthy interpreted his result as indicative of no-neutrino $\beta^-\beta^-$ decay, with half-life 2×10^{17} years.

This result of McCarthy is, however, completely contradicted by Awschalom's work on the same $\beta^-\beta^-$ decay. Awschalom (1956) used a 4π scintillation counter spectrometer similar to McCarthy's, and again recorded the numbers of e_1^-, e_2^- time-coincidences in the various intervals of $\epsilon_1+\epsilon_2$, with source and control alternately in the counting position. On the other hand, Awschalom's experiment was performed in a salt mine at a depth of 1100 feet, cutting down the cosmic ray background considerably. In fact, the majority of his residual background e_1^-, e_2^- time-coincidences were thought to arise from Compton electrons, due to low-energy γ-rays associated with the decay of $^{40}_{19}\mathrm{K}$ in the natural potassium of the mine. The residual background was further decreased by shielding the counters with one inch of iron.

The source used was a 182-mg sample of 84% $^{48}_{20}\mathrm{Ca}$ in the form of $CaCO_3$, and the control a 182-mg sample of chemically pure ${}_{20}\mathrm{Ca}$ of natural isotopic abundance, again in the form of $CaCO_3$. A 200-mg source of $^{96}_{40}\mathrm{Zr}$ in the form of ZrO_2 was also available. The data obtained are given in table 3.

It is evident that there is no statistically significant difference in the number of events with $^{48}_{20}\mathrm{Ca}$, or $^{96}_{40}\mathrm{Zr}$ on one hand, and natural ${}_{20}\mathrm{Ca}$ on the other. Consequently,

lower limits of 2×10^{18} years and 5×10^{17} years were assigned to the half-lives of the decays $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$, and $^{96}_{40}\mathrm{Zr} \rightarrow {}^{96}_{42}\mathrm{Mo}$, respectively.

Table 3

Sample	Running time (hours)	$e_1^- e_2^-$ coincidence counting rates (events per 100 hr) 2–4·5 MeV	3–4·75 MeV
$^{48}_{20}\mathrm{Ca}$	640	7·4	4·4
Natural $_{20}\mathrm{Ca}$	252	4·8	4·4
$^{96}_{40}\mathrm{Zr}$	504	5·5	4·9

Awschalom offered no explanation for the discrepancy between his result and that of McCarthy, but did point out that his experiment was done under more favourable circumstances, i.e. with a far lower background than McCarthy's (4·4 against 60 events per hundred hours). An experimental re-investigation of this $\beta^-\beta^-$ decay ($^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$) would thus appear to be in order.†

The most recent $\beta\beta$ experiment reported is the work of Cowan, Harrison, Langer and Reines (1956) on the $\beta^-\beta^-$ decay $^{150}_{60}\mathrm{Nd} \rightarrow {}^{150}_{62}\mathrm{Sm}$. The energy release, from the measured atomic masses of $^{150}_{60}\mathrm{Nd}$ and $^{150}_{62}\mathrm{Sm}$, is 3·7 MeV (see figure 9), which is the second highest available energy for $\beta\beta$ decay, the first being 4·3 MeV, released in the decay $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$. This, together with the fact that the Z of $^{150}_{62}\mathrm{Sm}$ is approximately three times that of $^{48}_{22}\mathrm{Ti}$, renders the estimated half-life of $^{150}_{60}\mathrm{Nd} \rightarrow {}^{150}_{62}\mathrm{Sm}$ shorter, by a factor of roughly two, than that for $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$ (see §§ 3, 4). In fact, the estimated half-lives for the $\beta^-\beta^-$ decay $^{150}_{60}\mathrm{Nd} \rightarrow {}^{150}_{62}\mathrm{Sm}$, with or without the emission of neutrinos, are shorter than the corresponding estimates for any other $\beta\beta$ decay process.

The source used in the Cowan, Harrison, Langer and Reines experiment was a 2·1 g sample of Nd_2O_3 prepared from $_{60}$Nd enriched to 88% $^{150}_{60}$Nd, and the control a similar sample of Nd_2O_3 prepared from $_{60}$Nd depleted to 0·06% $^{150}_{60}$Nd. Both source and control were mounted between layers of aluminium coated with Mylar film, immersed in a liquid scintillator, and viewed on each side, from a distance of 3 inches, by a photomultiplier tube. Light shields were provided and an axle permitted periodic interchange of the two samples. This system was placed in the centre of a 600-litre tank of liquid scintillator, the tank as a whole being viewed by twelve photomultiplier tubes operated in parallel; these twelve tubes provided an effective anti-coincidence shield for charged particle background. The apparatus was surrounded by a minimum of four inches of lead or bismuth on all sides.

In recording data, pairs of pulses associated with e_1^-, e_2^- time-coincident events in either the 'source' pair, or 'control' pair of photomultipliers, were accepted if they were not in coincidence with pulses from the guard photomultipliers. The conditions of acceptance were that each pulse in such a pair must lie within 0·35 and 8·0 MeV and that the two pulses must be coincident within 0·5 μsec. The energies of such pairs of pulses were summed electronically and analysed with a 100-channel pulse height analyser, the energy calibration being based on the peaks produced by minimum ionizing cosmic rays passing

† See footnote on p. 163.

through the scintillator. In this way the number of time-coincident events in the various $\epsilon_1+\epsilon_2$ intervals were recorded, both with the source and the control.

Data was collected for 358 hours. For no value of $\epsilon_1+\epsilon_2$ was there observed any statistically significant difference between the number of e_1^-, e_2^- time-coincidences with the source and with the control. Moreover, the distribution curve of e_1^-, e_2^- time-coincident events as a function of $\epsilon_1+\epsilon_2$ was flat between 3·5 and 5 Mev. A lower limit on the half-life of the $\beta^-\beta^-$ decay $^{150}_{60}\mathrm{Nd}\rightarrow{}^{150}_{62}\mathrm{Sm}$ was assigned as 4×10^{19} years.

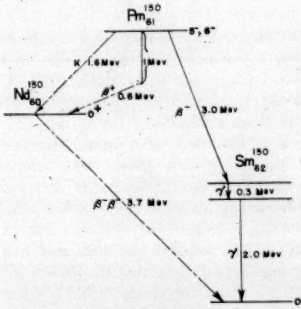


Figure 9. Energy level scheme of $\beta^-\beta^-$ decay : $^{150}_{60}\mathrm{Nd}\rightarrow{}^{150}_{62}\mathrm{Sm}$.

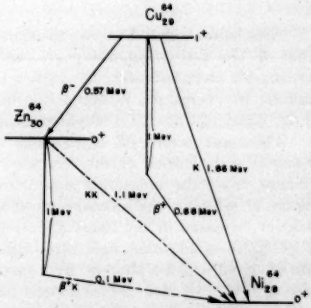


Figure 10. Energy level scheme of KK capture : $^{64}_{30}\mathrm{Zn}\rightarrow{}^{64}_{28}\mathrm{Ni}$.

The number of events in the flat distribution was about 280 in 358 hours for each 0·5 Mev step in the kinetic energy sum region $3{\cdot}5\ \mathrm{Mev}\leqslant\epsilon_1+\epsilon_2\leqslant5{\cdot}0\ \mathrm{Mev}$. Most of these events may be ascribed to the natural radio-activity of the photo-multiplier tubes viewing the source and the control, and to Compton electrons arising from gamma rays produced by cosmic-ray neutron capture in the detector shield.

One experiment on a $\beta\beta$ process where double negatron (K, K) capture is anticipated has been reported. Berthelot, Chaminade, Levi and Papineau (1953) studied the KK capture $^{64}_{30}\mathrm{Zn}\rightarrow{}^{64}_{28}\mathrm{Ni}$, which has a 1·1 Mev energy release (see figure 10), and attempted to count (though not in coincidence) the 8-kev x-ray photons emitted when the K-shell of $^{64}_{28}\mathrm{Ni}$ is filled. A sample of $_{30}\mathrm{Zn}$ containing 49% $^{64}_{30}\mathrm{Zn}$ was available as a source, and a proportional counter, followed by a 10-channel pulse height analyser, was used for the detection of the x-ray photons. No effect over background was observed and a lower limit of 10^{16} years was set on the half-life of the KK capture $^{64}_{30}\mathrm{Zn}\rightarrow{}^{64}_{28}\mathrm{Ni}$; this lower limit is sufficiently small to be consistent with the theoretical estimates for both the no-neutrino and the two-neutrino KK capture in $^{64}_{30}\mathrm{Zn}$.

§3. Theory of $\beta\beta$ Decay

3.1. *Nucleon–Lepton Interaction*

We shall now give a general discussion of the $\beta\beta$ decay process, taking into account the non-conservation of parity in the nucleon–lepton interaction (Lee

and Yang 1956, 1957, Lee 1957, Wu, Ambler, Hayward, Hoppes and Hudson 1957, Garwin, Lederman and Weinrich 1957, Friedman and Telegdi 1957). Prior to the discovery of the non-conservation of parity, the theory of $\beta\beta$ decay was developed by several authors (Mayer 1935, Racah 1937, Furry 1939, Touschek 1948, Tiomno 1950, Sliv 1950, Primakoff 1952, Konopinski 1955, Winter 1955, Rosen 1957, Blin-Stoyle and Rosen, unpublished). Some recent theoretical papers, dealing with the relation of the non-conservation of parity to the Majorana description of the neutrino and to the possible non-conservation of lepton charge, all being relevant to $\beta\beta$ decay, should also be noted (McLennan 1957, Case 1957, Pauli 1957, Enz 1957, Marx 1957).

We first consider the most general, Lorentz invariant, nucleon–lepton interaction Hamiltonian density H_{Dirac} (to be referred to as H_{D}), appropriate to the case $\nu \not\equiv \bar{\nu}$, in which the neutrino and the anti-neutrino are postulated to be distinguishable operationally (see § 1). We assume further that :

(*a*) The nucleon and lepton quantized field amplitudes entering into H_{D} are coupled in a local manner, so that H_{D} is invariant under the product of the transformations of space inversion (P), time reversal (T) and charge conjugation (C) (Lüders 1954, 1957, Lüders and Zumino 1957, Pauli 1955, Schwinger 1951, 1953, Bell 1955, Jost 1957).

(*b*) H_{D} need not conserve the total lepton charge. Then

$$H_{\text{D}} = \sum_{\lambda} (\psi_{\text{p}}^{\dagger}\, \Gamma^{(\lambda)}\, \psi_{\text{n}}) \{C_{\lambda}[\psi_{\text{e}}^{\dagger}\, \Gamma^{(\lambda)}\, (1+\delta_{\lambda}\gamma_5)\, \psi_{\nu}] + D_{\lambda}[\psi_{\text{e}}^{\dagger}\, \Gamma^{(\lambda)}\, (1+\eta_{\lambda}\gamma_5)\, \psi_{\bar{\nu}}]\} + \text{h.c.} \qquad \ldots\ldots(12)$$

In equation (12) Σ_{λ} covers the five Lorentz invariant interactions, i.e. λ is successively scalar (S), vector (V), tensor (T), axial vector (A), and pseudo-scalar (P); C_{λ}, D_{λ} and $C_{\lambda}\,\delta_{\lambda}$, $D_{\lambda}\eta_{\lambda}$ are respectively the parity conserving and parity non-conserving nucleon–lepton coupling constants. $\Gamma^{(\text{S})} = \gamma_4$, $\Gamma^{(\text{V})} = \gamma_4\gamma_j$, $\Gamma^{(\text{T})} = (2\sqrt{2}i)^{-1}\gamma_4(\gamma_j\gamma_k - \gamma_k\gamma_j)$, $\Gamma^{(\text{A})} = i^{-1}\gamma_4\gamma_j\gamma_5$, $\Gamma^{(\text{P})} = \gamma_4\gamma_5$, where the γ_j ($j = 1, 2, 3, 4$; summation over repeated subscripts j is implied, e.g. in equation (12)) are the four hermitian anticommuting Dirac matrices and $\gamma_5 = \gamma_1\gamma_2\gamma_3\gamma_4$.

$\psi_{\text{p}}, \psi_{\text{n}}, \psi_{\text{e}}, \psi_{\nu}, \psi_{\bar{\nu}}$ are quantized field amplitude operators which destroy the types of particle indicated by the subscripts and create the corresponding types of anti-particle. Thus, for example,

$$\psi_{\nu}(\mathbf{r}) = \sum_{\mathbf{k},\,s} \{\mathrm{d}_{\nu}(\mathbf{k},\kappa,s)\, u(\mathbf{k},\kappa,s;\,\mathbf{r}) + \mathrm{d}_{\bar{\nu}}^{\dagger}(\mathbf{k},\kappa,s)\, \mathscr{C}^{*}\, v^{*}(\mathbf{k},\kappa,s;\,\mathbf{r})\} = \mathscr{C}^{*}\, \tilde{\psi}_{\bar{\nu}}^{\dagger}(\mathbf{r}),$$

$$\psi_{\bar{\nu}}(\mathbf{r}) = \sum_{\mathbf{k},\,s} \mathrm{d}_{\bar{\nu}}(\mathbf{k},\kappa,s)\, v(\mathbf{k},\kappa,s;\,\mathbf{r}) + \mathrm{d}_{\nu}^{\dagger}(\mathbf{k},\kappa,s)\, \mathscr{C}^{*}\, u^{*}(\mathbf{k},\kappa,s;\,\mathbf{r}) = \mathscr{C}^{*}\, \tilde{\psi}_{\nu}^{\dagger}(\mathbf{r}). \qquad \ldots\ldots(13)$$

In equation (13) $\mathbf{k}, \kappa = |\mathbf{k}|$‡, s are momentum, energy and spin-orientation eigenvalues characterizing the four component spinor neutrino and anti-neutrino single-particle states $u(\mathbf{k},\kappa,s;\,\mathbf{r})$ and $v(\mathbf{k},\kappa,s;\,\mathbf{r})$. The superscripts *, † denote complex and hermitian conjugates respectively, and the superscript $\sim$ indicates the transpose of the spinors u, v, or of the matrix operator $\mathscr{C}$.

‡ The rest mass of the neutrino can be taken as zero. Empirically, it is found to be less than 1/500 of the electron mass (see Ridley 1956).

$\mathscr{C}$ is the Dirac charge conjugation matrix operator defined by

$$u(-\mathbf{k}, -\kappa, -s; \mathbf{r}) = \mathscr{C}^* v^*(\mathbf{k}, \kappa, s; \mathbf{r}),$$
$$v(-\mathbf{k}, -\kappa, -s; \mathbf{r}) = \mathscr{C}^* u^*(\mathbf{k}, \kappa, s; \mathbf{r}),$$

so that $\mathscr{C}^*\mathscr{C} = 1$.

$d_\nu^\dagger(\mathbf{k}, \kappa, s)$, $d_{\bar\nu}^\dagger(\mathbf{k}, \kappa, s)$ are the neutrino, anti-neutrino creation operators into the neutrino, anti-neutrino single-particle states $u(\mathbf{k}, \kappa, s; \mathbf{r})$, $v(\mathbf{k}, \kappa, s; \mathbf{r})$, while $d_\nu(\mathbf{k}, \kappa, s)$, $d_{\bar\nu}(\mathbf{k}, \kappa, s)$ are the corresponding destruction operators out of these states.

It is perhaps not superfluous to point out explicitly that if ψ_e, ψ_n, ψ_p in equation (12) are replaced by $(1+\delta_\lambda^{(e)}\gamma_5)\psi_e$, $(1+\delta_\lambda^{(n)}\gamma_5)\psi_n$, $(1+\delta_\lambda^{(p)}\gamma_5)\psi_p$ ($\delta_\lambda^{(e)}, \delta_\lambda^{(n)}, \delta_\lambda^{(p)}$ constants), the resultant H_D, with appropriate redefinitions of the coupling constants, is essentially the same as the original H_D of equation (12). This is a consequence of the fact that $\Gamma^{(\lambda)}$ and γ_5 either commute or anticommute and that $\gamma_5^2 = 1$.

It now follows from equation (12) that the necessary and sufficient condition on H_D for the conservation of lepton charge is

$$D_\lambda = 0 \quad (\lambda = S, \ldots, P), \qquad \ldots\ldots(14)$$

the combination of equations (12) and (14) yielding the nucleon–lepton interaction Hamiltonian density introduced by Lee and Yang (1956). When equation (14) is satisfied, H_D allows $n \to p + e^- + \bar\nu$ (equation (4*a*)) and $\nu + n \to p + e^-$ (equation (4*b*)), but forbids $n \to p + e^- + \nu$ (equation (5*a*)) and $\bar\nu + n \to p + e^-$ (equation (5*b*)), so that, as already noted in § 1, only two-neutrino $\beta\beta$ decay is possible.

In the context of H_D, the 'two-component neutrino' type coupling referred to in § 1, implies that either

$$\delta_\lambda = \eta_\lambda = 1 \quad (\lambda = S, \ldots, P) \qquad \ldots\ldots(15a)$$

or

$$\delta_\lambda = \eta_\lambda = -1 \quad (\lambda = S, \ldots, P). \qquad \ldots\ldots(15b)$$

Equations (15) are a generalization of the condition $\delta_\lambda = 1 (\lambda = S, \ldots, P)$ or $\delta_\lambda = -1 (\lambda = S, \ldots, P)$ for the 'two-component neutrino' type coupling, used in the H_D (equations (12), (14)) which conserves total lepton charge (Lee and Yang 1957, Landau 1957, Salam 1957). The 'two-component neutrino' type coupling of equations (15) (or equations (18), below) maximizes the various observable effects arising from the non-conservation of parity in the nucleon–lepton interaction (see discussion after equation (31)). Further, with this type of coupling, the 'interaction-induced' rest mass energy density of the neutrino field is proportional to

$$[\{(1 \pm \gamma_5)\psi_\nu\}^\dagger \gamma_4 \{(1 \pm \gamma_5)\psi_\nu\}] = [\psi_\nu^\dagger \gamma_4 (1 \mp \gamma_5)(1 \pm \gamma_5)\psi_\nu],$$

which is identically zero. Thus the assumption of a vanishing rest mass for the 'bare' neutrino here implies a vanishing rest mass for the 'physical' neutrino. Another consequence of this type of coupling is the vanishing of the magnetic moments of ν and of $\bar\nu$ ($\nu \not\equiv \bar\nu$), since any 'interaction-induced' magnetic moment orientation energy density, arising, for example, from the mechanism of the footnote on page 123, is here proportional to

$$[\{(1 \pm \gamma_5)\psi_\nu\}^\dagger \gamma_4 \{(\gamma_j\gamma_k - \gamma_k\gamma_j)/2i\} \{(1 \pm \gamma_5)\psi_\nu\}]$$

which is again identically zero.

It is worth noting that H_D is invariant under *each* of the transformations P, T, C, if the coupling constants satisfy the conditions of table 4. These conditions are completely analogous to those obtained by Lee, Oehme and Yang (1957) in the case where H_D conserves total lepton charge. As can be seen from table 4 and equations (15), it is an immediate consequence of 'two-component neutrino' type coupling that the corresponding H_D is *not* invariant under P, or under C, but that it *is* invariant under T (or PC).

Table 4

Transformation	Condition for invariance ($\lambda = S, \ldots, P$)	
P (or TC)	$\delta_\lambda = \eta_\lambda = 0$	
T (or PC)	$C_\lambda = C_\lambda^*$	$D_\lambda = D_\lambda^*$
	$\delta_\lambda = \delta_\lambda^*$	$\eta_\lambda = \eta_\lambda^*$
C (or PT)	$C_\lambda = C_\lambda^*$	$D_\lambda = D_\lambda^*$
	$\delta_\lambda = -\delta_\lambda^*$	$\eta_\lambda = -\eta_\lambda^*$

We next consider the most general, Lorentz invariant, nucleon–lepton interaction Hamiltonian density H_{Majorana} (H_M) appropriate to the case $\nu \equiv \bar{\nu}$, in which it is assumed that the neutrino and the anti-neutrino are indistinguishable operationally (and hence identical (see §1)). Here total lepton charge is not conserved, as, for example, in single β decay (see §1). Assuming further that H_M contains only local couplings and so is invariant under PTC, we may write

$$H_M = \sum_\lambda (\psi_p^\dagger \Gamma^{(\lambda)} \psi_n) F_\lambda (\psi_e^\dagger \Gamma^{(\lambda)} (1 + \epsilon_\lambda \gamma_5) \phi_\nu) + \text{h.c.} \qquad \ldots\ldots(16)$$

where

$$\sum_\lambda \ldots, \Gamma^{(\lambda)}, \gamma_5, \psi_p, \psi_n, \psi_e$$

are defined as in equation (12); F_λ and $F_\lambda \epsilon_\lambda$ are, respectively, parity conserving and parity non-conserving nucleon–lepton coupling constants and ϕ_ν is the quantized field amplitude operator which destroys and creates Majorana neutrinos ν, i.e.

$$\begin{aligned}\phi_\nu(\mathbf{r}) &= \sum_{\mathbf{k},s} \{\mathrm{d}_\nu(\mathbf{k},\kappa,s)\, u(\mathbf{k},\kappa,s;\mathbf{r}) + \mathrm{d}_\nu^\dagger(\mathbf{k},\kappa,s)\, \mathscr{C}^* u^*(\mathbf{k},\kappa,s;\mathbf{r})\} \\ &= \mathscr{C}^* \tilde{\phi}_\nu^\dagger(\mathbf{r}),\end{aligned} \qquad \ldots\ldots(17)$$

the $u(\mathbf{k},\kappa,s;\mathbf{r})$, $\mathscr{C}$, $\mathrm{d}_\nu(\mathbf{k},\kappa,s)$, $\mathrm{d}_\nu^\dagger(\mathbf{k},\kappa,s)$ being defined as in equation (13).

Analogously to the case of H_D, 'two-component neutrino' type coupling is given for H_M by

$$\epsilon_\lambda = 1 \quad (\lambda = S, \ldots, P) \qquad \ldots\ldots(18a)$$

or

$$\epsilon_\lambda = -1 \quad (\lambda = S, \ldots, P), \qquad \ldots\ldots(18b)$$

and the conditions for the invariance of H_M under *each* of the transformations P, T, C are as in table 5. It is seen from table 5 and equations (18) that an immediate consequence of 'two-component neutrino' type coupling is again the non-invariance of the corresponding H_M under P or under C, and its invariance under T (or PC).

In general, the use of H_M (equation (16)) to calculate transition probabilities of single β decay and $\beta\beta$ decay gives the same results as the use of H_D (equation (12)) provided that (*a*) the choice of coupling constants

$$C_\lambda = D_\lambda = (1/\sqrt{2})F_\lambda;\ \delta_\lambda = \eta_\lambda = \epsilon_\lambda \quad (\lambda = S, ..., P) \qquad .....(19)$$

is made, and that (*b*) in specifying, for the Dirac case, the relevant groups of initial and final states, there is no distinction drawn between ν and $\bar{\nu}$. An example

Table 5

Transformation	Conditions for invariance ($\lambda = S, ..., P$)	
P (or TC)	$\epsilon_\lambda = 0$	
T (or PC)	$F_\lambda = F_\lambda^*$	$\epsilon_\lambda = \epsilon_\lambda^*$
C (or PT)	$F_\lambda = F_\lambda^*$	$\epsilon_\lambda = -\epsilon_\lambda^*$

of provision (*b*) is a single β decay calculation of the transition probability for the emission of an e^- with given momentum and spin-orientation, accompanied by a 'neutrino' (also with given momentum and spin-orientation), which may be *either* a ν *or* a $\bar{\nu}$, the choice between them not being made. Consequently, we shall, in general, calculate the various single β decay and $\beta\beta$ decay theoretical formulae using H_D and obtain the corresponding formulae appropriate to H_M from equation (19).

3.2. *Single* $\beta^\mp$ *Decay*

The basic transitions induced in first order by the H_D of equation (12) are the direct and inverse single $\beta^\mp$ decays given in equations (4) and (5). Employing the usual time-dependent perturbation formalism, we may calculate from H_D the various single $\beta^\mp$ decay transition matrix elements, and so find the quantities of physical interest. In particular, we may obtain the probability distribution $P(\mathbf{p}, \mathbf{k})$ for emission of an $e^\mp$ with momentum $\mathbf{p}$, accompanied by a $\bar{\nu}$, or by a ν, with momentum $\mathbf{k}$. For an allowed single $\beta^\mp$ transition from a non-oriented parent nucleus, $P(\mathbf{p}, \mathbf{k})$ is given by

$$P(\mathbf{p}, \mathbf{k})\, d\epsilon\, d\Omega_{\mathbf{p}}\, d\Omega_{\mathbf{k}} = (2\pi)^{-5} \mathscr{F}(\pm Z, \epsilon)(p)(\epsilon+1)(\epsilon^{(0)} - \epsilon)^2 \times \xi(1 + b/(\epsilon+1) + a\mathbf{p}.\mathbf{k}/(\epsilon+1)k)\, d\epsilon\, d\Omega_{\mathbf{p}}\, d\Omega_{\mathbf{k}}. \qquad .....(20)$$

In equation (20), ϵ is the kinetic energy of the emitted $e^\mp$ (units being used throughout for which $c = 1$, $\hbar = 1$, $m_{\text{elec}} = 1$), and $\epsilon^{(0)}$ is the maximum kinetic energy, i.e. $\epsilon^{(0)}$ is the energy release of the single $\beta^\mp$ decay. $d\Omega_{\mathbf{p}}, d\Omega_{\mathbf{k}}$ are elements of solid angle associated with the directions of $\mathbf{p}$ and $\mathbf{k}$ respectively. $\mathscr{F}(\pm Z, \epsilon)$ is the Fermi factor, resulting from the effect of the Coulomb field of the daughter nucleus on the emitted $e^\mp$ (see, for example, Blatt and Weisskopf 1952). Very approximately,

$$\mathscr{F}(\pm Z, \epsilon) = \pm \left[\frac{2\pi Z}{137}\frac{\epsilon+1}{p}\right]\left[1 - \exp\left(\mp \frac{2\pi Z}{137}\frac{\epsilon+1}{p}\right)\right]^{-1}.$$

The quantities ξ, b, a are given as follows :

$$\begin{aligned}\xi = &|M_F|^2[|C_S|^2(1+|\delta_S|^2)+|C_V|^2(1+|\delta_V|^2)\\ &\quad +|D_S|^2(1+|\eta_S|^2)+|D_V|^2(1+|\eta_V|^2)]\\ &+|M_{GT}|^2[|C_T|^2(1+|\delta_T|^2)+|C_A|^2(1+|\delta_A|^2)\\ &\quad +|D_T|^2(1+|\eta_T|^2)+|D_A|^2(1+|\eta_A|^2)]\\ \to &|M_F|^2[|F_S|^2(1+|\epsilon_S|^2)+|F_V|^2(1+|\epsilon_V|^2)]\\ &\quad +|M_{GT}|^2[|F_T|^2(1+|\epsilon_T|^2)+|F_A|^2(1+|\epsilon_A|^2)], \qquad \ldots\ldots(21)\end{aligned}$$

$$\begin{aligned}\xi b = &\pm 2\mathscr{R}\{|M_F|^2[C_S C_V^*(1+\delta_S\delta_V^*)+D_S D_V^*(1+\eta_S\eta_V^*)]\\ &+|M_{GT}|^2[C_T C_A^*(1+\delta_T\delta_A^*)+D_T D_A^*(1+\eta_T\eta_A^*)]\}\\ \to &\pm 2\mathscr{R}\{|M_F|^2[F_S F_V^*(1+\epsilon_S\epsilon_V^*)]\\ &\quad +|M_{GT}|^2[F_T F_A^*(1+\epsilon_T\epsilon_A^*)]\} \qquad \ldots\ldots(22)\end{aligned}$$

$$\begin{aligned}\xi a = &|M_F|^2[|C_V|^2(1+|\delta_V|^2)-|C_S|^2(1+|\delta_S|^2)\\ &\quad +|D_V|^2(1+|\eta_V|^2)-|D_S|^2(1+|\eta_S|^2)]\\ &+\tfrac{1}{3}|M_{GT}|^2[|C_T|^2(1+|\delta_T|^2)-|C_A|^2(1+|\delta_A|^2)\\ &\quad +|D_T|^2(1+|\eta_T|^2)-|D_A|^2(1+|\eta_A|^2)]\\ \to &|M_F|^2[|F_V|^2(1+|\epsilon_V|^2)-|F_S|^2(1+|\epsilon_S|^2)]\\ &\quad +\tfrac{1}{3}|M_{GT}|^2[|F_T|^2(1+|\epsilon_T|^2)-|F_A|^2(1+|\epsilon_A|^2)], \qquad \ldots\ldots(23)\end{aligned}$$

equation (19) being used for the passage to the Majorana case ($\nu \equiv \bar{\nu}$). Correction terms of approximately $Z/137$ in ξa and $(Z/137)^2$ in ξb have been omitted. The quantities

$$|M_F|^2 = \left|\left\langle \Psi_f^* \left| \sum_{n=1}^{A} \tau_n^{\pm} \right| \Psi_i \right\rangle\right|^2 \qquad \ldots\ldots(24)$$

$$|M_{GT}|^2 = \left|\left\langle \Psi_f^* \left| \sum_{n=1}^{A} \tau_n^{\pm} \boldsymbol{\sigma}_n \right| \Psi_i \right\rangle\right|^2 \qquad \ldots\ldots(25)$$

of equations (21)–(23) are the square moduli of the single $\beta^{\mp}$ decay Fermi and Gamow–Teller nuclear matrix elements, averaged over the spin-orientations of the parent nucleus, and summed over those of the daughter (see Blatt and Weisskopf 1952). In M_F and M_{GT}, Ψ_i, Ψ_f are the configuration space wave functions characterizing the states of the parent and daughter nuclei respectively ; τ_n^+, τ_n^- are the (isobaric spin) neutron into proton, and proton into neutron transformation operators for the nth nucleon ; and $\boldsymbol{\sigma}_n$ is the Pauli spin operator for the nth nucleon.

The half-life $T_{1/2}$ for the single $\beta^{\mp}$ decay is given by

$$\begin{aligned}(\ln 2)/T_{1/2} &= \int P(\mathbf{p},\mathbf{k})\, d\epsilon\, d\Omega_{\mathbf{p}}\, d\Omega_{\mathbf{k}}\\ &= (2\pi^3)^{-1}\xi \int_0^{\epsilon^{(0)}} \mathscr{F}(\pm Z,\epsilon)(\underline{p})(\epsilon+1)(\epsilon^{(0)}-\epsilon)^2(1+b/(\epsilon+1))\, d\epsilon.\end{aligned} \qquad \ldots\ldots(26)$$

We may also calculate from H_D the probability distribution $P(\mathbf{p}, \mathbf{n})$ for the emission of an $e^{\mp}$ with momentum $\mathbf{p}$ and rest frame spin-orientation specified by a unit vector $\mathbf{n}$, accompanied by a ν or a $\bar{\nu}$ with any possible momentum and spin-orientation. We find, again for an allowed single $\beta^{\mp}$ decay from a non-oriented parent nucleus

$$P(\mathbf{p}, \mathbf{n})\, d\epsilon\, d\Omega_{\mathbf{p}} = (2\pi)^{-4} \mathscr{F}(\pm Z, \epsilon)(p)(\epsilon+1)(\epsilon^{(0)}-\epsilon)^2 \times \xi\left(1 + \frac{b}{(\epsilon+1)} + G\frac{\mathbf{p}.\mathbf{n}}{\epsilon+1}\right), \quad \ldots\ldots(27)$$

where

$$\xi G = \pm 2\mathscr{R}\{|M_F|^2[|C_S|^2\delta_S - |C_V|^2\delta_V + |D_S|^2\eta_S - |D_V|^2\eta_V] + |M_{GT}|^2[|C_T|^2\delta_T - |C_A|^2\delta_A + |D_T|^2\eta_T - |D_A|^2\eta_A]\}$$
$$\rightarrow \pm 2\mathscr{R}\{|M_F|^2[|F_S|^2\epsilon_S - |F_V|^2\epsilon_V] + |M_{GT}|^2[|F_T|^2\epsilon_T - |F_A|^2\epsilon_A]\}, \quad \ldots\ldots(28)$$

equation (19) being once more used for passage to the Majorana case and correction terms of approximately $Z/137$ in ξG being omitted. From equation (27), the $e^{\mp}$ spin-polarization in the direction $\mathbf{n}$, $\mathscr{P}_{e^{\mp}}(\mathbf{n})$, is given by

$$\mathscr{P}_{e^{\mp}}(\mathbf{n}) = \frac{P(\mathbf{p},\mathbf{n}) - P(\mathbf{p},-\mathbf{n})}{P(\mathbf{p},\mathbf{n}) + P(\mathbf{p},-\mathbf{n})} = \frac{[G/(\epsilon+1)]\,\mathbf{p}.\mathbf{n}}{1 + b/(\epsilon+1)} = G\frac{\mathbf{p}.\mathbf{n}}{\epsilon+1}, \quad \ldots\ldots(29)$$

the last equality being valid since b, the 'Fierz interference' term, is known to be quite close to zero (Sherr and Miller 1954). Thus the $e^{\mp}$ spin polarization in the longitudinal direction, $\mathbf{n} = \mathbf{p}/p$, becomes

$$\mathscr{P}_{e^{\mp}}(\mathbf{n} = \mathbf{p}/p) = G\,p/(\epsilon+1) \quad \ldots\ldots(30)$$

or, for fast $e^{\mp}$ with $p \simeq \epsilon + 1$,

$$\mathscr{P}_{\text{fast } e^{\mp}}(\mathbf{n} = \mathbf{p}/p) \simeq G. \quad \ldots\ldots(31)$$

It is to be emphasized that G, and therefore $\mathscr{P}_{e^{\mp}}$, vanishes in the parity conserving case $\delta_\lambda = \eta_\lambda = 0$ or $\epsilon_\lambda = 0 (\lambda = S, \ldots, P)$ (see tables 4, 5). However, G may vanish even when H_D or H_M contain terms without conservation of parity; for example, when $C_\lambda = D_\lambda, \delta_\lambda = -\eta_\lambda (\lambda = S, \ldots, P)$; or $D_\lambda = 0, \delta_\lambda = 1 (\lambda = S, \ldots, P)$, $C_S = C_V, C_T = C_A$; or $\epsilon_\lambda = 1 (\lambda = S, \ldots, P), F_S = F_V, F_T = F_A$.‡

At the time of writing, experimental evidence that $\mathscr{P}_{\text{fast } e^{\mp}}(\mathbf{n} = \mathbf{p}/p) \simeq \mp 1$ seems to be accumulating. Thus G may be taken as ∓ 1 and the parity conserving values for $\delta_\lambda, \eta_\lambda, \epsilon_\lambda$ completely rejected. It is, however, clear from equations (20)–(31) that the conditions $T_{1/2}$ given (i.e. ξ effectively given), a given, $b = 0$,

‡ In the case $C_\lambda = D_\lambda, \delta_\lambda = -\eta_\lambda (\lambda = S, \ldots, P)$, though H_D contains terms without conservation of parity, all *observable* effects of non-conservation of parity, e.g. the e^- spin polarization, nevertheless vanish if one detects e^- accompanied by *either* a $\bar{\nu}$, *or* a ν. (See also Kahana and Pursey (1957), Lüders (1958), who give as the general conditions for the absence of observable effects of non-conservation of parity if no experimental distinction is made between $\bar{\nu}$ and ν:

$$\{C_\mu C_\lambda^* (\delta_\mu + \delta_\lambda^*) + D_\mu D_\lambda^* (\eta_\mu + \eta_\lambda^*)\} = 0\,; \quad \{C_\mu D_\lambda (\delta_\mu - \eta_\lambda) - C_\lambda D_\mu (\delta_\lambda - \eta_\mu)\} = 0\,; \quad (\lambda, \mu = S, \ldots, P).$$

These conditions are satisfied for $C_\lambda = D_\lambda, \delta_\lambda = -\eta_\lambda (\lambda = S, \ldots, P)$.) If, however, one studies the various observable effects of non-conservation of parity *only* for those e^- which are accompanied by, say, $\bar{\nu}$ rather than ν, the terms in D_λ in equations (21)–(28) must be dropped and *non-vanishing* values are obtained. Since, in the present discussion, we are assuming operationally *distinguishable* $\bar{\nu}$ and ν, an experimental separation of $e^-, \bar{\nu}$ events from e^-, ν events is possible at least in principle.

$G = \mp 1$ can be satisfied by various different sets of values for the $C_\lambda, D_\lambda, \delta_\lambda, \eta_\lambda$, or the $F_\lambda, \epsilon_\lambda$, provided only that all of the $|\delta_\lambda|$ and $|\eta_\lambda|$ or all of the $|\epsilon_\lambda|$ are equal to 1. We shall discuss more fully certain sets of values for the coupling constants at a later point, but notice here that if 'the two-component neutrino' coupling (equation (15) or (18)) is used, G can be made equal to ∓ 1 by suitable choice of C_λ, D_λ or F_λ.

The formulae in equations (20)–(31) reduce, when $D_\lambda = 0 (\lambda = \mathrm{S}, \ldots, \mathrm{P})$, to the expressions for $P(\mathbf{p}, \mathbf{k})$, $P(\mathbf{p}, \mathbf{n})$, etc. previously derived by several authors (Jackson, Treiman and Wyld 1957a, b, Alder, Stech and Winther 1957, Ebel and Feldman 1957, Lewis 1957, Curtis and Lewis 1957) from the H_D (equations (12), (14)) which conserves total lepton charge. As can be seen from equations (20)–(31), identical results are obtained in single $\beta^\mp$ decay for the case $\nu \not\equiv \bar{\nu}$, with conservation of total lepton charge, and for the case $\nu \equiv \bar{\nu}$, the same numerical values for $C_\lambda, \delta_\lambda$ and $F_\lambda, \epsilon_\lambda$ being required for the correlation of a given body of experimental data.

3.3. *General Formulation of the $\beta^\mp\beta^\mp$ Decay Transition Probability Amplitude*

The transitions induced in the second order by H_D (equation (12)) and by H_M (equation (16)) are the no-neutrino and two-neutrino $\beta^\mp\beta^\mp$ decays given, respectively, in equations (6), (3) and (7). From the usual time-dependent perturbation formalism, the $\beta^\mp\beta^\mp$ decay transition probabilities are proportional to the square moduli of the corresponding $\beta^\mp\beta^\mp$ decay transition matrix elements,

$$\langle \beta^\mp\beta^\mp; (A, Z \pm 2) | (A, Z) \rangle,$$

which for H_D, are

$$\langle \beta^\mp\beta^\mp; (A, Z \pm 2) | (A, Z) \rangle = \sum_v \frac{\langle \chi_\mathrm{f}^* \Psi_\mathrm{f}^* | H_\mathrm{D} | \Psi_v \chi_v \rangle \langle \chi_v^* \Psi_v^* | H_\mathrm{D} | \Psi_\mathrm{i} \chi_\mathrm{i} \rangle}{E_v - E_\mathrm{i}}. \qquad \ldots\ldots (32)$$

In equation (32), $\Psi_\mathrm{i}, \Psi_\mathrm{f}, \Psi_v$ are the wave functions, in configuration space, of the states of the parent (A, Z), daughter $(A, Z \pm 2)$ and 'intermediate or virtual' $(A, Z \pm 1)$ nuclei, respectively. $\chi_\mathrm{i}, \chi_\mathrm{f}, \chi_v$ are the wave functions, in occupation-number space, of the initial, final and 'intermediate' states of the lepton field. Thus the state represented by χ_i contains no leptons and that represented by χ_v contains $\mathrm{e}_1^\mp$ and $\bar{\nu}_{1 \text{ or } 2}$ or $\nu_{1 \text{ or } 2}$, or $\mathrm{e}_2^\mp$ and $\bar{\nu}_{2 \text{ or } 1}$ or $\nu_{2 \text{ or } 1}$, each with specified momentum and spin-orientation. χ_f, on the other hand, may contain either $\mathrm{e}_1^\mp$ and $\mathrm{e}_2^\mp$ (no-neutrino $\beta^\mp\beta^\mp$ decay) or $\mathrm{e}_1^\mp$ and $\mathrm{e}_2^\mp$ plus one of the pairs $(\bar{\nu}_1, \bar{\nu}_2)$, $(\bar{\nu}_1, \nu_2)$, $(\nu_1, \bar{\nu}_2)$ or (ν_1, ν_2) (two-neutrino $\beta^\mp\beta^\mp$ decay), all the particles having specified momentum and spin-orientation. E_i is the energy of the initial state of the nucleon–lepton system, represented by $\Psi_\mathrm{i}\chi_\mathrm{i}$, so that $E_\mathrm{i} = W_\mathrm{i}$, with W_i the energy of the parent nucleus. E_v is the energy of the vth 'intermediate' state, represented by $\Psi_v\chi_v$, so that $E_v = W_v + (\epsilon_1 + 1) + |\mathbf{k}_{1 \text{ or } 2}|$ or $E_v = W_v + (\epsilon_2 + 1) + |\mathbf{k}_{2 \text{ or } 1}|$, W_v being the energy of the vth state of the 'intermediate' nucleus.

3.4. *No-neutrino $\beta^\mp\beta^\mp$ Decay*

We now evaluate $\langle \beta^\mp\beta^\mp; | A, Z \pm 2) | (A, Z) \rangle$ in equation (32), first considering no-neutrino $\beta^\mp\beta^\mp$ decay. Here the $\bar{\nu}$ or ν, 'virtually present' in the 'intermediate' state, is not present in the final state, and therefore the sum over v includes

contributions from all possible $\bar{\nu}$ *and* ν momenta and spin-orientations, as well as from all possible Ψ_v. Evaluating the 'single β decay' type transition matrix elements of H_D,

$$\langle \chi_\nu^* \Psi_v^* | H_D | \Psi_i \chi_1 \rangle \quad \text{and} \quad \langle \chi_{\bar{\nu}}^* \Psi_f^* | H_D | \Psi_v \chi_\nu \rangle,$$

and summing over the spin-orientations of $\bar{\nu}$ and ν, equation (32) becomes

$$\begin{aligned}
&\langle \beta^\mp \beta^\mp; (A, Z \pm 2) | (A, Z) \rangle \\
&= (1 - P_{12}) \sum_{\nu, \mathbf{k}; \mu, \lambda} \left\{ \left\langle \Psi_f^* \left| \sum_{n=1}^{A} \tau_n^{(\pm)} \Gamma_n^{(\mu)} \phi^*(\pm Z, \epsilon_2; \mathbf{p}_2 . \mathbf{r}_n) \exp(i\mathbf{k} . \mathbf{r}_n) \right| \Psi_v \right\rangle \right. \\
&\quad \times \frac{\left\langle \Psi_v^* \left| \sum_{m=1}^{A} \tau_m^{(\pm)} \Gamma_m^{(\lambda)} \phi^*(\pm Z, \epsilon_1; \mathbf{p}_1 . \mathbf{r}_m) \exp(-i\mathbf{k} . \mathbf{r}_m) \right| \Psi_i \right\rangle}{(W_v - W_1) + (\epsilon_1 + 1) + |\mathbf{k}|} \\
&\quad \left. \times \left(a^\dagger(2) \Gamma^{(\mu)} \left(\frac{1}{2} + \frac{1}{2} \frac{\boldsymbol{\alpha} . \mathbf{k}}{|\mathbf{k}|} \right) \mathscr{C}^\dagger [I_{\mu\lambda} + J_{\mu\lambda} \gamma_5] \bar{\Gamma}^{(\lambda)} a^*(1) \right) \right\}, \quad \ldots\ldots (33)
\end{aligned}$$

where quantities previously undefined are as follows: P_{12} is a permutation operator interchanging the roles of the emitted $e_1^\mp$ and $e_2^\mp$. Thus $(1 - P_{12})$ ensures that the right-hand side of equation (33) is anti-symmetric in the momenta and spin-orientations of the emitted $e_1^\mp$ and $e_2^\mp$. $\Gamma_n^{(8)} = (\gamma_4)_n$, etc., is the indicated Dirac matrix for the nth nucleon, $\mathbf{r}_n$ is the space coordinate of the nth nucleon.

$$w(\mathbf{p}_j, \epsilon_j, \mathrm{s}_j; \mathbf{r}_n) = a(\mathbf{p}_j, \epsilon_j, \mathrm{s}_j) \phi(\pm Z, \epsilon_j; \mathbf{p}_j . \mathbf{r}_n) \equiv a(j) \phi(\pm Z, \epsilon_j; \mathbf{p}_j . \mathbf{r}_n)$$

(with $j = 1, 2$) is the four-component spinor single-particle state of the $e_j^\mp$, characterized by the indicated values of (asymptotic) momentum, kinetic energy and spin-orientation. We have

$$\phi(0, \epsilon_j; \mathbf{p}_j . \mathbf{r}) = \exp(i\mathbf{p}_j . \mathbf{r}),$$

while

$$|\phi(\pm Z, \epsilon_j; 0)|^2 = \mathscr{F}(\pm Z, \epsilon_j),$$

is the Fermi factor of equation (20).

$$\frac{1}{2}\left(1 + \frac{\boldsymbol{\alpha} . \mathbf{k}}{k}\right) = \frac{1}{2}\left(1 - \gamma_5 \frac{\boldsymbol{\sigma} . \mathbf{k}}{k}\right)$$

is the projection operator on positive energy four-component spinor single-particle states of the virtual $\bar{\nu}$ or ν, $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ being the Dirac velocity and spin matrix operators, related to the γ_j by

$$\boldsymbol{\alpha} = i\gamma_4 \boldsymbol{\Upsilon}, \quad \boldsymbol{\sigma} = (2i)^{-1} (\boldsymbol{\Upsilon} \times \boldsymbol{\Upsilon}),$$

with $\boldsymbol{\Upsilon} = (\gamma_1, \gamma_2, \gamma_3)$.

The $I_{\mu\lambda}, J_{\mu\lambda}$ are certain combinations of the nucleon–lepton coupling constants with and without conservation of parity, viz.

$$\left.\begin{aligned}
I_{\mu\lambda} = I_{\lambda\mu} &= C_\mu D_\lambda (1 - \delta_\mu \eta_\lambda) + C_\lambda D_\mu (1 - \delta_\lambda \eta_\mu), \\
J_{\mu\lambda} = -J_{\lambda\mu} &= C_\mu D_\lambda (\delta_\mu - \eta_\lambda) \quad - C_\lambda D_\mu (\delta_\lambda - \eta_\mu),
\end{aligned}\right\} \quad \ldots\ldots (34a)$$

which, in the Majorana case ($\nu \equiv \bar{\nu}$) become

$$\left.\begin{aligned}
I_{\mu\lambda} = I_{\lambda\mu} &= F_\mu F_\lambda (1 - \epsilon_\mu \epsilon_\lambda), \\
J_{\mu\lambda} = -J_{\lambda\mu} &= F_\mu F_\lambda (\epsilon_\mu - \epsilon_\lambda).
\end{aligned}\right\} \quad \ldots\ldots (34b)$$

Pauli (1957) and Enz (1957) have demonstrated quite generally that $I_{\mu\lambda}, J_{\mu\lambda}$ are the only combinations of the coupling constants that may appear in no-neutrino $\beta\beta$ decay transition matrix elements, if the neutrino mass is zero. (See also Kahana and Pursey 1957).

The symmetry or antisymmetry of $I_{\mu\lambda}, J_{\mu\lambda}$ in μ, λ simplifies the expression for $P_{12}\sum_{v,\mathbf{k};\,\mu,\lambda}$ in equation (33). We obtain

$$\langle\beta^{\mp}\beta^{\mp};\,(A,Z\pm2)|(A,Z)\rangle$$

$$=\sum_{v,\mathbf{k};\,\mu,\lambda}\left[\left\{\frac{\left\langle\Psi_f^*\left|\sum_{n=1}^{A}\tau_n^{(\pm)}\Gamma_n^{(\mu)}\phi^*(\pm Z,\epsilon_2;\,\mathbf{p}_2.\mathbf{r}_n)\exp(i\mathbf{k}.\mathbf{r}_n)\right|\Psi_v^*\right\rangle\times\left\langle\Psi_v^*\left|\sum_{m=1}^{A}\tau_m^{(\pm)}\Gamma_m^{(\lambda)}\phi^*(\pm Z,\epsilon_1;\,\mathbf{p}_1.\mathbf{r}_m)\exp(-i\mathbf{k}.\mathbf{r}_m)\right|\Psi_i\right\rangle}{(W_v-W_i)+(\epsilon_1+1)+|\mathbf{k}|}\right.\right.$$

$$\left.\left.-\frac{\left\langle\Psi_f^*\left|\sum_{m=1}^{A}\tau_m^{(\pm)}\Gamma_m^{(\lambda)}\phi^*(\pm Z,\epsilon_1;\,\mathbf{p}_1.\mathbf{r}_m)\exp(i\mathbf{k}.\mathbf{r}_m)\right|\Psi_v^*\right\rangle\times\left\langle\Psi_v^*\left|\sum_{n=1}^{A}\tau_n^{(\pm)}\Gamma_n^{(\mu)}\phi^*(\pm Z,\epsilon_2;\,\mathbf{p}_2.\mathbf{r}_n)\exp(-i\mathbf{k}.\mathbf{r}_n)\right|\Psi_i\right\rangle}{(W_v-W_i)+(\epsilon_2+1)+|\mathbf{k}|}\right\}\right.$$

$$\left.\times\left(a^{\dagger}(2)\,\Gamma^{(\mu)}\left(\frac{1}{2}+\frac{1}{2}\frac{\boldsymbol{\alpha}.\mathbf{k}}{|\mathbf{k}|}\right)\mathscr{C}^{-1}[I_{\mu\lambda}+J_{\mu\lambda}\gamma_5]\,\tilde{\Gamma}^{(\lambda)}\,a^*(1)\right)\right]. \qquad \ldots\ldots(35)$$

The sum over v in equation (35) may be evaluated quite accurately by a closure method (Primakoff 1952, Rosen 1957). The virtual $\bar{\nu}$ or ν momenta which make the dominant contribution to the sum over $\mathbf{k}$ in equation (35), are, as is best seen from equations (37) below, of the order of the average reciprocal distance between a pair of neutrons or protons in the parent or daughter nuclei. Thus the corresponding $\bar{\nu}$ or ν energies, $|\mathbf{k}|$, are of the order of 35 MeV. On the other hand, estimates of the magnitude of the nucleon recoil upon emitting, or absorbing the 'virtual' $\bar{\nu}$ or ν, indicate that the 'intermediate' nucleus states which make the dominant contribution to the sum over v in equation (35) have W_v-W_i no greater than 10 MeV, and probably much smaller. Consequently, W_v-W_i may be neglected relative to $|\mathbf{k}|$ in the energy denominator. Therefore, summing over v by closure, we obtain

$$\langle\beta^{\mp}\beta^{\mp};\,(A,Z\pm2)|(A,Z)\rangle$$

$$\simeq\left\langle\Psi_f^*\left|\sum_{\mu,\lambda;\,m,n}\tau_n^{(\pm)}\tau_m^{(\pm)}\Gamma_n^{(\mu)}\Gamma_m^{(\lambda)}\phi^*(\pm Z,\epsilon_2;\,\mathbf{p}_2.\mathbf{r}_n)\right.\right.$$

$$\left.\left.\times\phi^*(\pm Z,\epsilon_1;\,\mathbf{p}_1.\mathbf{r}_m)\,(M_{mn}^{(\mu)(\lambda)}+N_{mn}^{(\mu)(\lambda)})\right|\Psi_i^*\right\rangle, \qquad \ldots\ldots(36)$$

where

$$M^{(\mu)(\lambda)}_{mn} = (a^\dagger(2)\,\Gamma^{(\mu)}\mathscr{C}^\dagger[I_{\mu\lambda}+J_{\mu\lambda}\gamma_5]\,\tilde{\Gamma}^{(\lambda)}\,a^*(1))$$
$$\times\left[\frac{1}{2}\int\frac{d\mathbf{k}}{(2\pi)^3}\exp(-i\mathbf{k}.\mathbf{r}_{mn})\frac{(\epsilon_2-\epsilon_1)}{(|\mathbf{k}|+(\epsilon_1+1)(|\mathbf{k}|+(\epsilon_2+1))}\right] \quad \ldots\ldots(37a)$$

and

$$N^{(\mu)(\lambda)}_{mn} = (a^\dagger(2)\,\Gamma^{(\mu)}\boldsymbol{\alpha}\mathscr{C}^\dagger[I_{\mu\lambda}+J_{\mu\lambda}\gamma_5]\,\tilde{\Gamma}^{(\lambda)}\,a^*(1))$$
$$\times\left[\frac{1}{2}\int\frac{d\mathbf{k}}{(2\pi)^3}\frac{\mathbf{k}}{k}\exp(-i\mathbf{k}.\mathbf{r}_{mn})\frac{2|\mathbf{k}|+(\epsilon_1+1)+(\epsilon_2+1)}{(|\mathbf{k}|+(\epsilon_1+1))(|\mathbf{k}|+(\epsilon_2+1))}\right]. \quad \ldots\ldots(37b)$$

We note that $M^{(\mu)(\lambda)}_{mn}$, $N^{(\mu)(\lambda)}_{mn}$ are, respectively, even, odd under inversion of the nucleon space coordinates ($\mathbf{r}_n\rightarrow-\mathbf{r}_n$; $\mathbf{r}_m\rightarrow-\mathbf{r}_m$).

Since the energy releases in $\beta^\mp\beta^\mp$ decay are never more than about 5 MeV, ϵ_1+1 and ϵ_2+1 are both small compared to $|\mathbf{k}|$. Using this to evaluate the integrals over $\mathbf{k}$ in equations (37), we have

$$M^{(\mu)(\lambda)}_{mn}\simeq\{a^\dagger(2)\,\Gamma^{(\mu)}\mathscr{C}^\dagger[I_{\mu\lambda}+J_{\mu\lambda}\gamma_5]\,\tilde{\Gamma}^{(\lambda)}a^*(1)\}\left[\frac{\pi^2}{(2\pi)^3 r_{mn}}(\epsilon_2-\epsilon_1)\right] \quad \ldots\ldots(38)$$

$$N^{(\mu)(\lambda)}_{mn}\simeq\{a^\dagger(2)\,\Gamma^{(\mu)}\boldsymbol{\alpha}\mathscr{C}^\dagger[I_{\mu\lambda}+J_{\mu\lambda}\gamma_5]\,\tilde{\Gamma}^{(\lambda)}a^*(1)\}\left[\frac{2\pi^2(-i\mathbf{r}_{mn})}{(2\pi)^3 r_{mn}{}^3}\right]. \quad \ldots\ldots(39)$$

Equations (36), (38) and (39) give

$$\langle\beta^\mp\beta^\mp;\,(A,Z\pm2)|(A,Z)\rangle$$

in the closure approximation.

Since the de Broglie wavelengths of the emitted electrons are large compared with nuclear dimensions, we have, provided $Z/137$ is not too great,

$$\phi(\pm Z,\epsilon;\,\mathbf{p}.\mathbf{r})\simeq\phi(\pm Z,\epsilon;\,0)\{1+i\mathbf{p}.\mathbf{r}\}. \quad \ldots\ldots(40)$$

Now, as seen in §2, all $\beta^\mp\beta^\mp$ decays of physical interest proceed between states Ψ_i, Ψ_f of the parent and daughter nuclei, each of which has zero spin and even parity. Therefore only that part of the operator sandwiched between Ψ_i and Ψ_f in equation (36) which is invariant under nucleon space and spin coordinate rotation and nucleon space coordinate inversion contributes to

$$\langle\beta^\mp\beta^\mp;\,(A,Z\pm2)|(A,Z)\rangle.$$

In the closure approximation, then, the matrix element for $0^+\rightarrow0^+$ no-neutrino $\beta^\mp\beta^\mp$ decay transitions, is, from equations (36) and (38)–(40),

$$\langle\beta^\mp\beta^\mp;\,(A,Z\pm2)|(A,Z)\rangle$$
$$\simeq\sum_{\mu,\lambda}\left\{\frac{1}{8\pi}\left(a^\dagger(2)\,\Gamma^{(\mu)}\left\{(\epsilon_2-\epsilon_1)+\frac{\boldsymbol{\alpha}}{3}.(\mathbf{p}_2-\mathbf{p}_1)\right\}\mathscr{C}^\dagger[I_{\mu\lambda}+J_{\mu\lambda}\gamma_5]\,\tilde{\Gamma}^{(\lambda)}a^*(1)\right)\right.$$
$$\left.\times\left\langle\Psi_f^*\left|\sum_{n,m}\frac{\tau_n(\pm)\,\tau_m(\pm)\,\Gamma_n{}^{(\mu)}\,\Gamma_m{}^{(\lambda)}}{r_{mn}}\right|\Psi_i\right\rangle\phi^*(\pm Z,\epsilon_2;\,0)\,\phi^*(\pm Z,\epsilon_1;\,0)\right\}. \quad \ldots\ldots(41)$$

Since the nucleons are non-relativistic, we may use in equation (41)

$$\Gamma_n^{(S)} \simeq \Gamma_n^{(V)} \simeq 1;\ \Gamma_n^{(T)} \simeq \Gamma_n^{(A)} \simeq \boldsymbol{\sigma}_n;\ \Gamma_n^{(P)} \simeq 0. \qquad \ldots\ldots(42a)$$

Thus, in equation (41),

$$\Gamma_n^{(S)}\Gamma_m^{(S)} \simeq \Gamma_n^{(V)}\Gamma_m^{(V)} \simeq \Gamma_n^{(S)}\Gamma_m^{(V)} \simeq \Gamma_n^{(V)}\Gamma_m^{(S)} \simeq 1_n 1_m;$$

$$\Gamma_n^{(T)}\Gamma_m^{(T)} \simeq \Gamma_n^{(A)}\Gamma_m^{(A)} \simeq \Gamma_n^{(T)}\Gamma_m^{(A)} \simeq \Gamma_n^{(A)}\Gamma_m^{(T)} \simeq \boldsymbol{\sigma}_n \cdot \boldsymbol{\sigma}_m;$$

$$\Gamma_n^{(\lambda)}\Gamma_m^{(P)} \simeq \Gamma_n^{(P)}\Gamma_m^{(\lambda)} \simeq 0 \quad (\lambda = S, \ldots, P). \qquad \ldots\ldots(42b)$$

From the square modulus of the transition matrix element in equation (41), using also equation (42*b*), we obtain, by summing over the spin-orientations of $e_1^{\mp}, e_2^{\mp}$, the probability distribution $P(\mathbf{p}_1, \mathbf{p}_2)$ for emission of the $e_1^{\mp}, e_2^{\mp}$ with momenta $\mathbf{p}_1, \mathbf{p}_2$. With θ the angle between $\mathbf{p}_1$ and $\mathbf{p}_2$,

$$\begin{aligned}
&P(\mathbf{p}_1, \mathbf{p}_2)\, 2\pi \sin\theta\, d\theta\, d\epsilon_1\, d\epsilon_2 \\
&\quad = (2\pi)^{-5}(p_1)(\epsilon_1+1)(p_2)(\epsilon_2+1)\,\delta(\epsilon^{(0)} - \epsilon_1 - \epsilon_2)\,\mathscr{F}(\pm Z, \epsilon_1) \\
&\qquad \times \mathscr{F}(\pm Z, \epsilon_2)\,[A(\epsilon_1, \epsilon_2) + B(\epsilon_1, \epsilon_2)\cos\theta]\,\tfrac{1}{2}\sin\theta\, d\theta\, d\epsilon_1\, d\epsilon_2 \\
&\quad \simeq (2\pi)^{-5}\left(\frac{2\pi Z}{137}\right)^2 [1 - \exp(\mp 2\pi Z/137)]^{-2}(\epsilon_1+1)^2(\epsilon_2+1)^2\,\delta(\epsilon^{(0)} - \epsilon_1 - \epsilon_2) \\
&\qquad \times [A(\epsilon_1, \epsilon_2) + B(\epsilon_1, \epsilon_2)\cos\theta]\,\tfrac{1}{2}\sin\theta\, d\theta\, d\epsilon_1\, d\epsilon_2, \qquad \ldots\ldots(43)
\end{aligned}$$

where

$$\begin{aligned}
&[A(\epsilon_1, \epsilon_2) + B(\epsilon_1, \epsilon_2)\cos\theta] \\
&\equiv \Bigg\{ \left|\frac{I_{SS} + I_{VV}}{2}\right|^2 f_1 + |I_{SV}|^2 f_2 + |J_{SV}|^2 f_3 + \left|\frac{I_{SS} - I_{VV}}{2}\right|^2 f_4 \\
&\quad + \frac{2}{3}\mathscr{R}\left[\left(\frac{I_{SS} + I_{VV}}{2}\right) I_{SV}^{\,*} f_5 + \left(\frac{I_{SS} + I_{VV}}{2}\right)\left(\frac{I_{SS} - I_{VV}}{2}\right)^* f_6 + \left(\frac{I_{SS} - I_{VV}}{2}\right)^* I_{SV} f_7\right]\Bigg\} \\
&\quad \times \left|\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} / r_{nm} \right| \Psi_i \right\rangle\right|^2 + \Bigg\{ \left|\frac{I_{TT} + I_{AA}}{2}\right|^2 f_1 + |I_{TA}|^2 f_2 + |J_{TA}|^2 f_3 \\
&\quad + \left|\frac{I_{TT} - I_{AA}}{2}\right|^2 f_4 + \frac{2}{9}\mathscr{R}\left[\frac{1}{3}\left(\frac{I_{TT} + I_{AA}}{2}\right) I^*_{TA} f_5 - \left(\frac{I_{TT} + I_{AA}}{2}\right)\left(\frac{I_{TT} - I_{AA}}{2}\right)^* f_6\right. \\
&\quad \left. + \left(\frac{I_{TT} - I_{AA}}{2}\right)^* I_{TA} f_7\right]\Bigg\} \left|\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} \boldsymbol{\sigma}_n \cdot \boldsymbol{\sigma}_m / r_{nm} \right| \Psi_i \right\rangle\right|^2 \\
&\quad + 2\mathscr{R}\Bigg\{\left[\left(\frac{I_{SS} + I_{VV}}{2}\right)\left(\frac{I_{TT} + I_{AA}}{2}\right)^* f_1 + I_{SV} I_{TA}^{\,*} f_2 + J_{SV} J_{TA}^{\,*} f_3\right. \\
&\quad - \left(\frac{I_{SS} - I_{VV}}{2}\right)\left(\frac{I_{TT} - I_{AA}}{2}\right)^* f_4 + \frac{1}{9}\left(\left(\frac{I_{SS} + I_{VV}}{2}\right) I_{TA}^{\,*} + \left(\frac{I_{TT} + I_{AA}}{2}\right) I_{SV}^{\,*}\right) f_5 \\
&\quad + \left(\frac{1}{3}\left(\frac{I_{SS} - I_{VV}}{2}\right)\left(\frac{I_{TT} + I_{AA}}{2}\right)^* - \left(\frac{I_{SS} + I_{VV}}{2}\right)\left(\frac{I_{TT} - I_{AA}}{2}\right)^*\right) f_6 + \left(\frac{1}{9}\left(\frac{I_{SS} - I_{VV}}{2}\right) I_{TA}^{\,*}\right. \\
&\quad \left.\left. - \frac{1}{3}\left(\frac{I_{TT} - I_{AA}}{2}\right) I_{SV}^{\,*}\right) f_7\right] \left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} / r_{nm} \right| \Psi_i \right\rangle \\
&\qquad \times \left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} \boldsymbol{\sigma}_n \cdot \boldsymbol{\sigma}_m / r_{mn} \right| \Psi_i \right\rangle^*\Bigg\}. \qquad \ldots\ldots(44)
\end{aligned}$$

with

$$f_1 = \frac{1}{9H_1H_2}[(p_1^2+p_2^2)(H_1H_2+1)-2p_1^2p_2^2+(p_1^2+p_2^2-2(H_1H_2+1))\,\mathbf{p}_1\cdot\mathbf{p}_2]$$

$$f_2 = \frac{1}{9H_1H_2}[(p_1^2+p_2^2)(H_1H_2+1)+2p_1^2p_2^2-(p_1^2+p_2^2+2(H_1H_2+1))\,\mathbf{p}_1\cdot\mathbf{p}_2]$$

$$f_3 = \frac{1}{9H_1H_2}[(p_1^2+p_2^2)(H_1H_2-1)+2p_1^2p_2^2-(p_1^2+p_2^2+2(H_1H_2-1))\,\mathbf{p}_1\cdot\mathbf{p}_2]$$

$$f_4 = \frac{1}{H_1H_2}(H_1-H_2)^2(H_1H_2-1+\mathbf{p}_1\cdot\mathbf{p}_2)$$

$$f_5 = \frac{1}{3}\left(\frac{1}{H_1}+\frac{1}{H_2}\right)(p_1^2+p_2^2-2\mathbf{p}_1\cdot\mathbf{p}_2)$$

$$f_6 = \left(\frac{1}{H_1}-\frac{1}{H_2}\right)(\mathbf{p}_1-\mathbf{p}_2)\cdot(\mathbf{p}_1H_2+\mathbf{p}_2H_1)$$

$$f_7 = \left(\frac{1}{H_1}-\frac{1}{H_2}\right)(p_1^2-p_2^2), \qquad \ldots\ldots(45)$$

and $H_1 \equiv \epsilon_1+1, H_2 \equiv \epsilon_2+1$. Since each f_i is of the form $f_i'(\epsilon_1,\epsilon_2)+f_i''(\epsilon_1,\epsilon_2)\cos\theta$, the form $A(\epsilon_1,\epsilon_2)+B(\epsilon_1,\epsilon_2)\cos\theta$ for the left-hand side of equation (44) is justified. The δ function of equation (43) expresses the fact that in a no-neutrino $\beta^\mp\beta^\mp$ decay, the sum $\epsilon_1+\epsilon_2$ of the kinetic energies of the electrons is the same in each and every disintegration, being equal to the decay's energy release $\epsilon^{(0)}$ (see §§ 1, 2).

From equations (43)–(45) we can calculate the $e_1^\mp, e_2^\mp$ angular correlation function $C(\epsilon_1,\epsilon^{(0)}-\epsilon_1;\theta)\frac{1}{2}\sin\theta d\theta$; the probability distribution in kinetic energy of either of the emitted electrons, say $e_1^\mp$, $P(\epsilon_1)$; and the half-life of the no-neutrino $\beta^\mp\beta^\mp$ decay, $T_{1/2}$, viz.

$$C(\epsilon_1,\epsilon^{(0)}-\epsilon_1;\theta) = 1+\left[\frac{B(\epsilon_1,\epsilon^{(0)}-\epsilon_1)}{A(\epsilon_1,\epsilon^{(0)}-\epsilon_1)}\right]\cos\theta, \qquad \ldots\ldots(46)$$

$$P(\epsilon_1) = \int_0^{\epsilon^{(0)}}\int_0^{\pi} P(\mathbf{p}_1,\mathbf{p}_2)\,2\pi\sin\theta\,d\theta\,d\epsilon_2$$

$$\simeq (2\pi)^{-5}\left(\frac{2\pi Z}{137}\right)^2\left(1-\exp\left(\mp\frac{2\pi Z}{137}\right)\right)^{-2}(\epsilon_1+1)^2(\epsilon^{(0)}-\epsilon_1+1)^2A(\epsilon_1,\epsilon^{(0)}-\epsilon_1), \qquad \ldots\ldots(47)$$

$$\frac{\ln 2}{T_{1/2}} = \int_0^{\epsilon^{(0)}} P(\epsilon_1)\,d\epsilon_1 \simeq (2\pi)^{-5}\left(\frac{2\pi Z}{137}\right)^2\left(1-\exp\left(\mp\frac{2\pi Z}{137}\right)\right)^{-2}$$

$$\times\int_0^{\epsilon^{(0)}}(\epsilon_1+1)^2(\epsilon^{(0)}-\epsilon_1+1)^2A(\epsilon_1,\epsilon^{(0)}-\epsilon_1)\,d\epsilon_1. \qquad \ldots\ldots(48)$$

To estimate $T_{1/2}$ from equations (48) we first consider the nuclear matrix elements appearing in $A(\epsilon_1,\epsilon_2)$ (see equations (44), (45)), taking

$$\left|\left\langle\Psi_f^*\left|\sum_{m,n}\frac{\tau_m^{(\pm)}\tau_n^{(\pm)}}{r_{mn}}\right|\Psi_i\right\rangle\right|^2 \simeq \frac{1}{100}\frac{1}{R^2}, \qquad \ldots\ldots(49a)$$

$$\left|\left\langle \Psi_f^* \left| \sum_{m,n} \frac{\tau_m^{(\pm)}\tau_n^{(\pm)}}{r_{mn}} \boldsymbol{\sigma}_m . \boldsymbol{\sigma}_n \right| \Psi_i \right\rangle\right|^2 \simeq \frac{1}{100}\frac{1}{R^2}, \qquad \ldots\ldots(49b)$$

where $R = 1{\cdot}2 \times 10^{-13} A^{1/3}$ cm is the nuclear radius. The estimates of equations (49*a*), (49*b*) are similar to those of Konopinski (1955) and are based on the replacement of each r_{mn} by R and on the assumption that

$$\begin{aligned}
&\left|\left\langle \Psi_f^* \left| \sum_{m,n} \tau_m^{(\pm)}\tau_n^{(\pm)} \right| \Psi_i \right\rangle\right| \\
&\quad = \left| \sum_v \left\langle \Psi_f^* \left| \sum_{m=1}^{A} \tau_m^{(\pm)} \right| \Psi_v \right\rangle \left\langle \Psi_v^* \left| \sum_{n=1}^{A} \tau_n^{(\pm)} \right| \Psi_i \right\rangle \right| \\
&\quad \simeq \sum_v \left| \left\langle \Psi_v^* \left| \sum_{m=1}^{A} \tau_m^{(\pm)} \right| \Psi_i \right\rangle \right|^2 \simeq \mathcal{N}_v' \left| \left\langle \Psi_v^* \left| \sum_m \tau_m^{(\pm)} \right| \Psi_i \right\rangle \right|^2, \dagger
\end{aligned} \qquad \ldots\ldots(49c)$$

$$\begin{aligned}
&\left|\left\langle \Psi_f^* \left| \sum_{m,n} \tau_m^{(\pm)}\tau_n^{(\pm)} \boldsymbol{\sigma}_m . \boldsymbol{\sigma}_n \right| \Psi_i \right\rangle\right| \\
&\quad = \left| \sum_v \left\langle \Psi_f^* \left| \sum_m \tau_m^{(\pm)} \boldsymbol{\sigma}_m \right| \Psi_v \right\rangle . \left\langle \Psi_v^* \left| \sum_n \tau_n^{(\pm)} \boldsymbol{\sigma}_n \right| \Psi_i \right\rangle \right| \\
&\quad \simeq \sum_v \left| \left\langle \Psi_v^* \left| \sum_m \tau_m^{(\pm)} \boldsymbol{\sigma}_m \right| \Psi_i \right\rangle \right|^2 \simeq \mathcal{N}_v'' \left| \left\langle \Psi_v^* \left| \sum_m \tau^{(\pm)}{}_m \boldsymbol{\sigma}_m \right| \Psi_i \right\rangle \right|^2,
\end{aligned} \qquad \ldots\ldots(49d)$$

$\mathcal{N}_v', \mathcal{N}_v''$ being the number of states Ψ_v of the 'intermediate' nucleus with appreciable single $\beta^{\mp}$ decay type nuclear matrix elements

$$\left\langle \Psi_v^* \left| \sum_m \tau_m^{(\pm)} \right| \Psi_i \right\rangle, \quad \left\langle \Psi_v^* \left| \sum_m \tau_m^{(\pm)} \boldsymbol{\sigma}_m \right| \Psi_i \right\rangle.$$

If we take the square moduli of these to be roughly the same as for allowed, unfavoured, single $\beta^{\mp}$ transitions, viz. $\simeq 1/100$, and assume $\mathcal{N}_v' \simeq \mathcal{N}_v'' \simeq 10$, equations (49*c*) and (49*d*) yield the estimates of equations (49*a*), (49*b*). The numerical values used for the allowed unfavoured single $\beta^{\mp}$ decay nuclear matrix elements are indicated by the range of comparative half-lives among the single $\beta^{\mp}$ unstable nuclei of medium and large A.

It should now be noted that, for 'two-component neutrino' type coupling, whether in the case $\tilde{\nu} \not\equiv \nu$ (equations (15*a*), (15*b*)) or in the case $\tilde{\nu} \equiv \nu$ (equations (18*a*), (18*b*)), all the combinations $I_{\mu\lambda}, J_{\mu\lambda}$ of nucleon–lepton coupling constants (equations (34*a*), (34*b*)) vanish, *whatever* the values of C_λ, D_λ or F_λ. Consequently $\langle \beta^{\mp}\beta^{\mp}; (A, Z \pm 2) | (A, Z) \rangle$, which is linear in $I_{\mu\lambda}, J_{\mu\lambda}$ (equations (33), (35)), or $P(\mathbf{p}_1, \mathbf{p}_2)$ which is bilinear in $I_{\mu\lambda}, J_{\mu\lambda}$ (equations (43), (44)), also vanish and no-neutrino $\beta^{\mp}\beta^{\mp}$ decay cannot occur, even if $\tilde{\nu}$ is identical to ν. This absence of the no-neutrino decay, when 'two-component neutrino' coupling is used (see § 1), has been pointed out in the Majorana ($\tilde{\nu} \equiv \nu$) case by several authors (Case 1957,

† The first approximate equality in equation (49*c*), and hence the whole estimate of this equation, holds only if no isobaric spin selection rule operates to make $\langle \Psi_f^* | \sum_{m,n} \tau_m^{(\pm)} \tau_n^{(\pm)} | \Psi_i \rangle = 0$. See the footnote in connection with equation (72*a*), p. 158.

McLennan 1957, Pauli 1957). Physically it arises from the fact that with 'two-component neutrino' type coupling, the 'neutrino' (be it $\tilde{\nu}$ or ν if $\tilde{\nu} \equiv \nu$, or ν if $\tilde{\nu} \equiv \nu$) can only be emitted by a neutron into one spin-orientation state (say spin parallel to momentum) and only absorbed by a neutron if it occupies the opposite spin-orientation state (spin antiparallel to momentum). Obviously, then, no-neutrino $\beta^{\mp}\beta^{\mp}$ decay (as in equation (6) or equation (3)) cannot occur, since the 'neutrino' emitted by one neutron is in the wrong spin state for absorption by the other.†

Consider now a choice of coupling constants which (a) is consistent with most of the available single $\beta^{\mp}$ decay data, and (b) gives a non-vanishing rate for no-neutrino $\beta^{\mp}\beta^{\mp}$ decay.

These two requirements immediately rule out both the parity conserving and 'two-component neutrino' values for δ_λ, η_λ or ϵ_λ. In what follows, we shall, for simplicity, adopt the Majorana point of view, $\tilde{\nu} \equiv \nu$, and take in equation (16):

$$\left.\begin{aligned} &\epsilon_{\mathrm{V}} = \epsilon_{\mathrm{A}} = 1, \quad \epsilon_{\mathrm{S}} = \epsilon_{\mathrm{T}} = -1, \\ &F_{\mathrm{V}} \neq 0, F_{\mathrm{A}} \neq 0, \quad \text{with} \quad |F_{\mathrm{A}}| \simeq |F_{\mathrm{V}}|, \\ &F_{\mathrm{T}} = 0, F_{\mathrm{S}} \neq 0, \quad \text{but} \quad |F_{\mathrm{S}}/F_{\mathrm{V}}|^2 \ll 1. \end{aligned}\right\} \qquad \ldots\ldots(50)$$

Note that if F_{S}, as well as F_{T}, were zero, we would effectively have 'two-component neutrino' type coupling since all the non-vanishing F_λ would then have the same ϵ_λ ($\epsilon_\lambda = 1, \lambda = \mathrm{V}, \mathrm{A}$). Thus with the choice of coupling constants given in equation (50), the rate of no-neutrino $\beta\beta$ decay will be expected to vanish as $|F_{\mathrm{S}}/F_{\mathrm{V}}|^2$ does.

Substituting the coupling constants of equation (50) into equations (21)–(23) and (28), we have

$$\left.\begin{aligned} &\xi = |M_{\mathrm{F}}|^2 [2|F_{\mathrm{V}}|^2 (1 + |F_{\mathrm{S}}/F_{\mathrm{V}}|^2)] + |M_{\mathrm{GT}}|^2 [2|F_{\mathrm{A}}|^2], \\ &b = 0 : G = \mp 1, \\ &a = \begin{cases} \dfrac{(1 - |F_{\mathrm{S}}/F_{\mathrm{V}}|^2)}{(1 + |F_{\mathrm{S}}/F_{\mathrm{V}}|^2)} & \text{pure Fermi transition,} \\ -1/3 & \text{pure Gamow–Teller transition.} \end{cases} \end{aligned}\right\} \qquad \ldots\ldots(51)$$

As already noted, the values of b and G (Fierz interference coefficient and fast $e^{\mp}$ longitudinal spin polarization, respectively) given in equation (51) are in agreement with available experimental data. The values of a (electron–neutrino angular correlation coefficient) in equation (51), for $|F_{\mathrm{S}}/F_{\mathrm{V}}|^2 \leqslant 1/3$, agree with the experiments of Hermannsfeldt, Maxson, Stahelin and Allen (1957), Maxson, Allen and Jentschke (1955), Good and Lauer (1957), Alford and Hamilton (1957); however, $a = -\frac{1}{3}$ for a pure Gamow–Teller transition is in disagreement with the

† It may be remarked that in the 'twin neutrino' theory of Mayer and Telegdi (1957) and of Preston (1957), the nucleon–lepton coupling constants correspond (in our notation) e.g. to $C_{\mathrm{V}} \neq 0$, $D_{\mathrm{A}} \neq 0$, $\delta_{\mathrm{V}} = -1$, $\eta_{\mathrm{A}} = 1$, all other C_λ, D_λ being zero. As a result the only non-vanishing $I_{\mu\lambda}, J_{\mu\lambda}$ are $I_{\mathrm{VA}}, J_{\mathrm{VA}}$, neither of which enters into equation (44); thus, as noted by these authors, though for a different reason from our own, no-neutrino $\beta\beta$ decay does not occur. In papers by Cavanagh, Coleman, Ridley and Turner (1958) and by Gatto (1958) one possibility considered is (again in our notation) $C_{\mathrm{S}} \neq 0$, $C_{\mathrm{T}} \neq 0$, $\delta_{\mathrm{S}} = \delta_{\mathrm{T}} = -1$; $D_{\mathrm{V}} \neq 0$, $D_{\mathrm{A}} \neq 0$, $\eta_{\mathrm{V}} = \eta_{\mathrm{A}} = 1$; all other C_λ, D_λ being zero. In this case, $I_{\mathrm{TA}}, J_{\mathrm{TA}}$; $J_{\mathrm{SV}}, J_{\mathrm{SV}}$ are different from zero, so that, from equations (44) and (48), no-neutrino $\beta\beta$ decay may be expected to occur.

original, though not the most recent, interpretation of the experiment of Rustad and Ruby (1955). An experiment by Goldhaber, Grodzins and Sunyar (1958) on the value P_ν of the longitudinal spin polarization of neutrinos emitted in pure Gamow–Teller K-capture, gives $P_\nu \simeq -1$; since it can be shown that

$$P_\nu = -2\mathscr{R}[(F_T \epsilon_T + F_A \epsilon_A)(F_T^* + F_A^*)]/[|F_T + F_A|^2 + |F_T \epsilon_T + F_A \epsilon_A|^2],$$

we may conclude that $\epsilon_T = \epsilon_A = 1$. This result, with equations (28) and (21) for G and the empirical value $G \simeq \mp 1$, implies that $F_A \neq 0, F_T = 0$, as in equation (50).†

With the coupling constants of equation (50), we have from equation (34*b*),

$$\left.\begin{aligned} I_{SS} = I_{VV} = I_{TT} = I_{AA} = 0, \\ I_{SV} = 2F_S F_V, \qquad I_{TA} = 0, \\ J_{SV} = -2F_S F_V, \quad J_{TA} = 0. \end{aligned}\right\} \qquad \ldots\ldots(52)$$

Thus, from equation (44),

$$A(\epsilon_1, \epsilon_2) = 4|F_S|^2 |F_V|^2 \left(\frac{2[(p_1^2 + p_2^2) H_1 H_2 + 2p_1^2 p_2^2]}{9H_1 H_2}\right) \times \left| \left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} / r_{nm} \right| \Psi_i \right\rangle \right|^2$$

$$\simeq 4|F_S|^2 |F_V|^2 \left(\frac{2[(p_1^2 + p_2^2) H_1 H_2 + 2p_1^2 p_2^2]}{9H_1 H_2}\right) \left[\frac{1}{100} \frac{1}{R^2}\right], \qquad \ldots\ldots(53a)$$

$$\frac{B(\epsilon_1, \epsilon_2)}{A(\epsilon_1, \epsilon_2)} = -\frac{(p_1^2 + p_2^2 + 2H_1 H_2) p_1 p_2}{[(p_1^2 + p_2^2) H_1 H_2 + 2p_1^2 p_2^2]}. \qquad \ldots\ldots(53b)$$

Using equations (53*a*), (48) and equations (53*b*), (46) to obtain the half-life for no-neutrino $\beta^\mp \beta^\mp$ decay and the $e_1^\mp, e_2^\mp$ angular correlation function, we have :

$$T_{1/2} \simeq 3 \times 10^{14 \pm 2} \left|\frac{F_V}{F_S}\right|^2 \left(\frac{60}{Z}\right)^2 \left[1 - \exp\left(\mp \frac{2\pi Z}{137}\right)\right]^2 \left(\frac{A}{150}\right)^{20/3} \left(\frac{8}{\epsilon^{(0)}}\right)^6 \text{yrs}$$

$$\simeq 1 \times 10^{15 \pm 2} \left(\frac{60}{Z}\right)^2 \left[1 - \exp\left(\mp \frac{2\pi Z}{137}\right)\right]^2 \left(\frac{A}{150}\right)^{20/3} \left(\frac{8}{\epsilon^{(0)}}\right)^6 \text{yrs}, \qquad \ldots\ldots(54)$$

$$C(\epsilon_1, \epsilon^{(0)} - \epsilon_1; \theta) = 1 - \left\{ \left[\frac{(p_1^2 + p_2^2 + 2H_1 H_2) p_1 p_2}{(p_1^2 + p_2^2) H_1 H_2 + 2p_1^2 p_2^2} \right]_{\epsilon_2 = \epsilon^{(0)} - \epsilon_1} \right\} \cos\theta. \qquad \ldots\ldots(55)$$

In equation (54) $\epsilon^{(0)}$ is the energy release of the (no-neutrino) $\beta^\mp \beta^\mp$ decay, defined in §2 and expressed in units of $m_{elec} c^2$; $2|F_V|^2$, determined from ξ of equation (51), is approximately 10^{-23} ; ξ. in turn, is determined from equation (26), using observed half-lives of pure Fermi single $\beta^\mp$ decay transitions of the allowed, favoured type, for which $|M_F|^2 = 2$, $|M_{GT}|^2 = 0$ (Feenberg 1955). The quantity $|F_V/F_S|^2$ has been set equal to its lower limit, 3 (see discussion after equation (51)). The proportionality of $1/T_{1/2}$ to $(\epsilon^{(0)})^6$ is a very rough approximation to the actual dependence of the integral in equation (48) on its upper limit.

† *Note added in proof* : Very recent experiments by Herrmannsfeldt, Burman, Stahelin, Allen and Braid (1958) and by Burgy, Krohn, Novey, Ringo and Telegdi (1958) yield further convincing evidence that $|F_T/F_A| \ll 1$, with the latter experiment also demonstrating that $F_A \simeq -F_V$.

The ± 2 in the exponent is an estimate of the overall error, which arises largely from the uncertainty in the estimates of the nuclear matrix elements (see equations (49)). Notice that the uncertainty in $|F_V/F_S|^2$ is not included in the error quoted, and that if F_S (as well as F_T) is shown to be zero, then $T_{1/2}$ will be infinite.

We conclude this discussion by emphasizing that the vanishing of $I_{\mu\lambda}$, $J_{\mu\lambda}$ in equations (34) and the resultant absence of no-neutrino $\beta^\mp\beta^\mp$ decay can be explained on the basis of theories which do, or which do not, conserve total lepton charge: in the first case the appropriate interaction Hamiltonian density is H_D (equation (12): $\nu \equiv \tilde{\nu}$) with $D_\lambda = 0$ ($\lambda = S, \ldots, P$); and in the second, it is either $H_D(D_\lambda \neq 0)$, or H_M (equation (16): $\nu \equiv \tilde{\nu}$), both with 'two-component neutrino' type coupling.

3.5. *Two-neutrino $\beta^\mp\beta^\mp$ Decay*

When all the $I_{\mu\lambda}$ and $J_{\mu\lambda}$ of equations (34) vanish, then the *only* form of $\beta\beta$ decay that can occur involves the emission of two neutrinos. As just noted, these quantities are, in fact, zero in two extreme and aesthetically attractive cases:

(*a*) $\nu \equiv \tilde{\nu}$ so that total lepton charge is not conserved but with 'two-component neutrino' type coupling (equations (18)) assumed to be valid. The corresponding nucleon–lepton interaction Hamiltonian density is then, from equation (16),

$$H_M = \sum_\lambda (\psi_p^\dagger \Gamma^{(\lambda)} \psi_n) F_\lambda (\psi_e^\dagger \Gamma^{(\lambda)} (1 \pm \gamma_5) \phi_\nu) + \text{h.c.} \qquad \ldots\ldots(56)$$

(*b*) $\nu \equiv \tilde{\nu}$, with conservation of total lepton charge (equation (14)). Assuming the validity of 'two-component neutrino' coupling (equations (15)) in this case also the interaction Hamiltonian density is from equation (12)

$$H_D = \sum_\lambda (\psi_p^\dagger \Gamma^{(\lambda)} \psi_n) C_\lambda (\psi_e^\dagger \Gamma^{(\lambda)} (1 \pm \gamma_5) \psi_\nu) + \text{h.c.} \qquad \ldots\ldots(57)$$

The interaction Hamiltonian densities of equations (56) and (57) predict identical results in experiments where the particle or anti-particle nature of the 'neutrino' is not studied explicitly; for example (I) single $\beta^\mp$ decay, (II) no-neutrino $\beta^\mp\beta^\mp$ decay, and (III) two-neutrino $\beta^\mp\beta^\mp$ decay. In more detail:

(I)
$$\left.\begin{aligned} n &\to p + e^- + \nu \\ p &\to n + e^+ + \nu \end{aligned}\right\}$$

in case (*a*) and

$$\left.\begin{aligned} n &\to p + e^- + \tilde{\nu} \\ p &\to n + e^+ + \nu \end{aligned}\right\}$$

in case (*b*). Here the square modulus of the single $\beta^\mp$ decay transition matrix element, $|\langle \beta^\mp; (A, Z \pm 1) | (A, Z) \rangle|^2$, is the same function of the F_λ in one case as it is of the C_λ in the other. This can be seen from equations (13), (17) for ϕ_ν and ψ_ν, or from equations (20)–(31) for various observable single $\beta^\mp$ effects.

(II)
$$\left.\begin{aligned} n_1 + n_2 &\to p_1 + p_2 + e_1^- + e_2^- \\ p_1 + p_2 &\to n_1 + n_2 + e_1^+ + e_2^+ \end{aligned}\right\}$$

in case (*a*) and case (*b*), and, as shown in §3.4, $|\langle\beta^{\mp}\beta^{\mp};(A,Z\pm2)|(A,Z)\rangle|^2$ vanishes for both cases.

(III)
$$\left.\begin{aligned} n_1+n_2 &\rightarrow p_1+p_2+e_1^-+e_2^-+\nu_1+\nu_2 \\ p_1+p_2 &\rightarrow n_1+n_2+e_1^++e_2^++\nu_1+\nu_2 \end{aligned}\right\}$$

for case (*a*), and

$$\left.\begin{aligned} n_1+n_2 &\rightarrow p_1+p_2+e_1^-+e_2^-+\bar{\nu}_1+\bar{\nu}_2 \\ p_1+p_2 &\rightarrow n_1+n_2+e_1^++e_2^++\nu_1+\nu_2 \end{aligned}\right\}$$

for case (*b*).

Since $\langle\beta^{\mp}\beta^{\mp};(A,Z\pm2)|(A,Z)\rangle$ is expressed essentially as the sum of products of $\langle\beta^{\mp};(A,Z\pm1)|(A,Z)\rangle$ (equation (32)), its square modulus is the same function of F_λ in (a) as of C_λ in (b). Physically, both cases give the same results for (III) because there is no re-absorption here, in a second step, of a 'virtual' neutrino, emitted in a first step.

We note that, with the 'old' parity conserving nucleon–lepton interaction, the Hamiltonian densities of equations (56) and (57) would become

$$H_{\mathrm{M}}=\sum_{\lambda}(\psi_{\mathrm{p}}^{\dagger}\Gamma^{(\lambda)}\psi_{\mathrm{n}})F_{\lambda}(\psi_{\mathrm{e}}^{\dagger}\Gamma^{(\lambda)}\phi_{\nu})+\mathrm{h.c.} \qquad \ldots\ldots(58)$$

(see equation (16), table 5), and

$$H_{\mathrm{D}}=\sum_{\lambda}(\psi_{\mathrm{p}}^{\dagger}\Gamma^{(\lambda)}\psi_{\mathrm{n}})C_{\lambda}(\psi_{\mathrm{e}}^{\dagger}\Gamma^{(\lambda)}\psi_{\nu})+\mathrm{h.c.} \qquad \ldots\ldots(59)$$

(see equations (12), (14) and table 4). These also predict identical results for (I) and (III), for precisely the same reasons as given above, in the case of the Hamiltonian densities of equations (56) and (57), but for (II), no-neutrino $\beta^{\mp}\beta^{\mp}$ decay, their predictions are entirely different. For the H_{D} of equation (59) all the D_λ are zero and therefore the $I_{\mu\lambda}, J_{\mu\lambda}$ (equation (34*a*)) vanish, whereas for the H_{M} of equation (58), $I_{\lambda\lambda}=F_\lambda^2$ (equation (34*b*)), and this is different from zero for at least one λ. Consequently, while the H_{D} of equation (59) gives zero probability for no-neutrino $\beta^{\mp}\beta^{\mp}$ decay, the H_{M} of equation (58) gives a non-vanishing rate for the process. Comparing this situation with that for H_{D} and H_{M} of equations (56) and (57), we see that the non-conservation of parity has destroyed the possibility of a decision in favour of $\nu\not\equiv\bar{\nu}$ with conservation of total lepton charge and against $\nu\equiv\bar{\nu}$, merely on the basis that two-neutrino, rather than no-neutrino, $\beta^{\mp}\beta^{\mp}$ decay is observed.

We proceed to an explicit calculation of $\langle\beta^{\mp}\beta^{\mp};(A,Z\pm2)|(A,Z)\rangle$ and the half-life for two-neutrino $\beta^{\mp}\beta^{\mp}$ decay using H_{M} of equation (56). As noted above, the results using H_{D} of equation (57) are obtained by replacing F_λ with C_λ in the expressions derived. We also notice that for *half-lives*, even the Hamiltonian densities of equations (58) and (59) give results very similar to those of equations (56) and (57).

We first replace, in equation (32), H_{D} by H_{M} of equation (56) and evaluate the 'single $\beta^{\mp}$ decay' type transition matrix elements, $\langle\chi_\nu^*\Psi_\nu^*|H_{\mathrm{M}}|\Psi_1\chi_1\rangle$ and

$\langle \chi_f^* \Psi_f^* | H_M | \Psi_v \chi_v \rangle$, that occur. We obtain

$$\langle \beta^{\mp}\beta^{\mp}; (A, Z\pm 2) | (A, Z)\rangle$$
$$= [1 - P_{12}^{(a)} - P_{12}^{(b)} + P_{12}^{(a)} P_{12}^{(b)}] \sum_{v,\mu,\lambda} O_{11}^{(\mu)(\lambda)}(v)$$
$$\times \frac{(a^{\dagger}(2)\, F_{\mu}\, \Gamma^{(\mu)}(1\pm\gamma_5)\mathscr{C}^* b^*(2))\,(a^{\dagger}(1)\, F_{\lambda}\, \Gamma^{(\lambda)}(1\pm\gamma_5)\mathscr{C}^* b^*(1))}{(W_v - W_i) + (\epsilon_1 + 1) + |\mathbf{k}_1|}, \qquad \ldots\ldots(60)$$

the notation being the same as that in equation (33) with the additional definitions

$$O_{11}^{(\mu)(\lambda)}(v) \equiv \left\langle \Psi_f^* \middle| \sum_{n=1}^{A} \tau_n^{(\pm)}\, \Gamma_n^{(\mu)}\, \phi^*(\pm Z, \epsilon_2; \mathbf{p}_2 . \mathbf{r}_n) \exp(-i\mathbf{k}_2 . \mathbf{r}_n) \middle| \Psi_v \right\rangle$$
$$\times \left\langle \Psi_v^* \middle| \sum_{m=1}^{A} \tau_m^{(\pm)}\, \Gamma_m^{(\lambda)}\, \phi^*(\pm Z, \epsilon_1; \mathbf{p}_1 . \mathbf{r}_m) \exp(-i\mathbf{k}_1 . \mathbf{r}_m) \middle| \Psi_1 \right\rangle \qquad \ldots\ldots(61)$$

$$u(\mathbf{k}_j, |\mathbf{k}_j|, s_j; \mathbf{r}) = b(\mathbf{k}_j, |\mathbf{k}_j|, s_j) \exp(i\mathbf{k}_j . \mathbf{r})$$
$$\equiv b(j) \exp(i\mathbf{k}_j . \mathbf{r}) \quad (j = 1, 2). \qquad \ldots\ldots(62)$$

$P_{12}^{(a)}$, $P_{12}^{(b)}$ are permutation operators for the emitted $e_1^{\mp}, e_2^{\mp}$ pair and the emitted ν_1, ν_2 pair. Their use ensures the antisymmetry of the $\beta^{\mp}\beta^{\mp}$ decay transition matrix element in equation (60) in the electron and neutrino momenta and spin-orientations.

We recall that the nucleons are non-relativistic, that the de Broglie wavelengths of the emitted leptons are large compared with nuclear dimensions (equations (40) and (42)), and that rather in the same way as in the discussion of §3.4 (see equations (50), (51) *et seq.*) the choice of coupling constants

$$|F_V| = |F_A| = F \neq 0; \quad F_S = F_T = 0; \quad \epsilon_V = \epsilon_A = 1 \qquad \ldots\ldots(63a)$$

offers a sufficiently accurate account of available single $\beta^{\mp}$ decay experiments. This choice of coupling constants is besides attractive from a general theoretical viewpoint (Sudarshan and Marshak 1957, 1958, Feynman and Gell-Mann 1958, Sakurai 1958) since the V–A interaction with 'two-component neutrino' type coupling is equivalent to an arbitrary nucleon–lepton interaction with 'two-component' type coupling for all of the fermions involved:

$$\sum_{\lambda} [\{(1+\gamma_5)\psi_p\}^{\dagger}\, \Gamma^{(\lambda)}\, \{(1+\gamma_5)\psi_n\}]\, G_{\lambda}\, [\{(1+\gamma_5)\psi_e\}^{\dagger}\, \Gamma^{(\lambda)}\, \{(1+\gamma_5)\phi_\nu\}] + \text{h.c.}$$
$$= 4(G_V - G_A)\, \{(\psi_p^{\dagger}\, \Gamma^{(V)}\, \psi_n)\, (\psi_e^{\dagger}\, \Gamma^{(V)}\, (1+\gamma_5)\, \phi_\nu)$$
$$- (\psi_p^{\dagger}\, \Gamma^{(A)}\, \psi_n)\, (\psi_e^{\dagger}\, \Gamma^{(A)}\, (1+\gamma_5)\, \phi_\nu)\} + \text{h.c.}$$
$$\equiv F\{(\psi_p^{\dagger}\, \Gamma^{(V)}\, \psi_n)\, (\psi_e^{\dagger}\, \Gamma^{(V)}\, (1+\gamma_5)\, \phi_\nu)$$
$$- (\psi_p^{\dagger}\, \Gamma^{(A)}\, \psi_n)\, (\psi_e^{\dagger}\, \Gamma^{(A)}\, (1+\gamma_5)\, \phi_\nu)\} + \text{h.c.} \qquad \ldots\ldots(63b)$$

We can now express those $O_{11}^{(\mu)(\lambda)}(v)$ of equation (61) which are associated

with non-vanishing F_μ, F_λ as :

$$O_{11}^{(\mathrm{V})(\mathrm{V})}(v) = \phi^*(\pm Z, \epsilon_2; 0)\phi^*(\pm Z, \epsilon_1; 0)\left\langle \Psi_f^* \middle| \sum_{n=1}^{A} \tau_n^{(\pm)} \middle| \Psi_v \right\rangle \times \left\langle \Psi_v^* \middle| \sum_{m=1}^{A} \tau_m^{(\pm)} \middle| \Psi_1 \right\rangle$$

$$O_{11}^{(\mathrm{A})(\mathrm{A})}(v) = \phi^*(\pm Z, \epsilon_2; 0)\phi^*(\pm Z, \epsilon_1; 0)\left\langle \Psi_f^* \middle| \sum_{n=1}^{A} \tau_n^{(\pm)} \boldsymbol{\sigma}_n \middle| \Psi_v \right\rangle \times \left\langle \Psi_v^* \middle| \sum_{m=1}^{A} \tau_m^{(\pm)} \boldsymbol{\sigma}_m \middle| \Psi_1 \right\rangle$$

$$O_{11}^{(\mathrm{V})(\mathrm{A})}(v) = \phi^*(\pm Z, \epsilon_2; 0)\phi^*(\pm Z, \epsilon_1; 0)\left\langle \Psi_f^* \middle| \sum_{n=1}^{A} \tau_n^{(\pm)} \middle| \Psi_v \right\rangle \times \left\langle \Psi_v^* \middle| \sum_{m=1}^{A} \tau_m^{\pm} \boldsymbol{\sigma}_m \middle| \Psi_1 \right\rangle = 0,$$

$$O_{11}^{(\mathrm{A})(\mathrm{V})}(v) = \phi^*(\pm Z, \epsilon_2; 0)\phi^*(\pm Z, \epsilon_1; 0)\left\langle \Psi_f^* \middle| \sum_{n=1}^{A} \tau_n^{(\pm)} \boldsymbol{\sigma}_n \middle| \Psi_v \right\rangle \times \left\langle \Psi_v^* \middle| \sum_{m=1}^{A} \tau_m^{(\pm)} \middle| \Psi_1 \right\rangle = 0, \qquad \ldots\ldots(63c)$$

the third and fourth of the equations (63*c*) being a consequence of the fact that the states Ψ_1, Ψ_f have zero spin and even parity, 0^+ (§§3.4 and 2) and of the further fact that for non-vanishing $\langle \Psi_v^* | \tau^{\pm} | \Psi_1 \rangle$ or $\langle \Psi_f^* | \tau^{\pm} | \Psi_v \rangle$, Ψ_v must be 0^+, while for non-vanishing $\langle \Psi^* | \tau^{\pm} \boldsymbol{\sigma} | \Psi_v \rangle$ or $\langle \Psi_v^* | \tau^{\pm} \boldsymbol{\sigma} | \Psi_1 \rangle$, Ψ_v must be 1^+. In equation (63*c*) the subscript 11 may be dropped since, within the approximation of these equations, $P_{12}^{(a)} O_{11}^{(\mu)(\lambda)}(v) = O_{11}^{(\mu)(\lambda)}(v)$ and similarly with $P_{12}^{(b)}$.

The main contribution to the sum over v in equation (60) comes, in general, from energetically low lying states Ψ_v of the 'intermediate' nucleus ; in fact, it may happen that either the lowest lying 0^+, or the lowest lying 1^+, state Ψ_v makes the dominant contribution. In any case, we may replace $W_v - W_1$ in the energy denominator of equation (60) by a suitable average value $\langle W_v - W_1 \rangle$, independent of v, and no more than a few MEV. The sum over v can then be performed by closure, and we obtain from equations (60), (61) and (63)

$$\begin{aligned}
&\langle \beta^{\mp}\beta^{\mp}; (A, Z \pm 2) | (A, Z) \rangle \\
&= [1 - P_{12}^{(a)} - P_{12}^{(b)} + P_{12}^{(a)} P_{12}^{(b)}] \left[\frac{F^2}{\langle W_v - W_1 \rangle + (\epsilon_1 + 1) + |\mathbf{k}_1|} \right] \\
&\qquad \times [\phi^*(\pm Z, \epsilon_2; 0)\phi^*(\pm Z, \epsilon_1; 0)] \\
&\times \Big\{ [a^\dagger(2)(1+\gamma_5)\mathscr{C}^* b^*(2)][a^\dagger(1)(1+\gamma_5)\mathscr{C}^* b^*(1)] \\
&\qquad \times \left\langle \Psi_f^* \middle| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} \middle| \Psi_1 \right\rangle \\
&+ \tfrac{1}{3}[a^\dagger(2)\boldsymbol{\sigma}(1+\gamma_5)\mathscr{C}^* b^*(2)][a^\dagger(1)\boldsymbol{\sigma}(1+\gamma_5)\mathscr{C}^* b^*(1)] \\
&\qquad \times \left\langle \Psi_f^* \middle| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} \boldsymbol{\sigma}_n \cdot \boldsymbol{\sigma}_m \middle| \Psi_1 \right\rangle \Big\} \qquad \ldots\ldots(64)
\end{aligned}$$

for the two-neutrino $\beta^{\mp}\beta^{\mp}$ decay matrix element for $0^+ \to 0^+$ transitions in the closure approximation.

We sum the square modulus of the transition matrix element in equation (64) over the spin-orientations of the $e_1^{\mp}, e_2^{\mp}, \nu_1, \nu_2$, average it over the neutrino momentum directions ($\mathbf{k}_i/|\mathbf{k}_i|$ for $i = 1, 2$), and call the resulting expression $[|\langle\beta^{\mp}\beta^{\mp}; (A, Z\pm 2)|(A, Z)\rangle|^2]_{\text{av}}$. With θ as the angle between $\mathbf{p}_1$ and $\mathbf{p}_2$, and $H_i \equiv \epsilon_i + 1 \,(i = 1, 2)$, we have

$$
\begin{aligned}
[|\langle\beta^{\mp}\beta^{\mp}; (A, Z\pm 2)|(A, Z)\rangle|^2]_{\text{av}} &= 2^4 |F|^4 \mathscr{F}(\pm Z, \epsilon_1)\mathscr{F}(\pm Z, \epsilon_2) \\
&\times \left\{[(K^2+L^2) - KL(1+\mathbf{p}_1\cdot\mathbf{p}_2/H_1H_2)]\left|\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)}\tau_m^{(\pm)} \right| \Psi_i \right\rangle\right|^2 \right. \\
&+ 3\left[(K^2+L)^2\left(1 - \frac{2}{3}\frac{\mathbf{p}_1\cdot\mathbf{p}_2}{H_1H_2}\right) + KL\left(1 - \frac{5}{3}\frac{\mathbf{p}_1\cdot\mathbf{p}_2}{H_1H_2}\right)\right] \\
&\times \left|\frac{1}{3}\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)}\tau_m^{(\pm)}\boldsymbol{\sigma}_n\cdot\boldsymbol{\sigma}_m \right| \Psi_i \right\rangle\right|^2 \\
&- 3KL\left(1 - \frac{1}{3}\frac{\mathbf{p}_1\cdot\mathbf{p}_2}{H_1H_2}\right) 2\mathscr{R}\left\{\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)}\tau_m^{(\pm)} \right| \Psi_i \right\rangle \right. \\
&\left.\left. \times \frac{1}{3}\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)}\tau_m^{(\pm)}\boldsymbol{\sigma}_n\cdot\boldsymbol{\sigma}_m \right| \Psi_i \right\rangle \right\}\right\} \\
&\equiv [A'(\epsilon_1, \epsilon_2, k_1, k_2) + B'(\epsilon_1, \epsilon_2, k_1, k_2)\cos\theta]\mathscr{F}(\pm Z, \epsilon_1)\mathscr{F}(\pm Z, \epsilon_2), \qquad \ldots\ldots(65)
\end{aligned}
$$

where

$$
\left.
\begin{aligned}
K &\equiv \frac{1}{\langle W_v - W_1\rangle + H_1 + |\mathbf{k}_1|} + \frac{1}{\langle W_v - W_1\rangle + H_2 + |\mathbf{k}_2|} \\
L &\equiv \frac{1}{\langle W_v - W_1\rangle + H_1 + |\mathbf{k}_2|} + \frac{1}{\langle W_v - W_1\rangle + H_2 + |\mathbf{k}_1|}
\end{aligned}
\right\}. \qquad \ldots\ldots(66)
$$

From equations (65) and (66), the probability distribution $Q(\mathbf{p}_1, \mathbf{p}_2)$ for the emission of $e_1^{\mp}, e_2^{\mp}$ with momenta $\mathbf{p}_1, \mathbf{p}_2$ is, with $\epsilon^{(0)}$ the energy release of the $\beta^{\mp}\beta^{\mp}$ decay,

$$
\begin{aligned}
&Q(\mathbf{p}_1, \mathbf{p}_2)\, 2\pi \sin\theta\, d\theta\, d\epsilon_1\, d\epsilon_2 \\
&= 2^{-5}\pi^{-7}(p_1)(\epsilon_1+1)(p_2)(\epsilon_2+1)\mathscr{F}(\pm Z, \epsilon_1)\mathscr{F}(\pm Z, \epsilon_2) \\
&\times \left[\int_0^{\epsilon^{(0)}}\int_0^{\epsilon^{(0)}} (A'(\epsilon_1, \epsilon_2, k_1, k_2) + B'(\epsilon_1, \epsilon_2, k_1, k_2)\cos\theta)\,\delta(\epsilon^{(0)} - \epsilon_1 - \epsilon_2 - k_1 - k_2)\right. \\
&\qquad \left.\times k_1^2\, dk_1\, k_2^2\, dk_2\right]\left[\frac{1}{2}\sin\theta\, d\theta\, d\epsilon_1\, d\epsilon_2\right] \\
&\simeq N(\epsilon_1+1)^2(\epsilon_2+1)^2\left[\int_0^{\epsilon^{(0)}-\epsilon_1-\epsilon_2} (A'(\epsilon_1, \epsilon_2, k_1, \epsilon^{(0)} - \epsilon_1 - \epsilon_2 - k_1)\right. \\
&\qquad \left. + B'(\epsilon_1, \epsilon_2, k_1, \epsilon^{(0)} - \epsilon_1 - \epsilon_2 - k_1)\cos\theta)\, k_1^2 (\epsilon^{(0)} - \epsilon_1 - \epsilon_2 - k_1)^2\, dk_1\right] \\
&\qquad \times \left[\frac{1}{2}\sin\theta\, d\theta\, d\epsilon_1\, d\epsilon_2\right] \qquad \ldots\ldots(67a)
\end{aligned}
$$

where
$$N \equiv 2^{-5}\pi^{-7}\left(\frac{2\pi Z}{137}\right)^2\left(1-\exp\left(\mp\frac{2\pi Z}{137}\right)\right)^2. \qquad \ldots\ldots,(67b)$$

From equation (67), with $A'(\epsilon_1, \epsilon_2, k_1, k_2)$ and $B'(\epsilon_1, \epsilon_2, k_1, k_2)$ defined by equations (65) and (66), we can calculate the $e_1^{\mp}, e_2^{\mp}$ angular correlation function, $C'(\epsilon_1, \epsilon_2, \theta)$; the probability distribution in kinetic energy of *either* of the emitted electrons, say $e_1^{\mp}$, $P'(\epsilon_1)$; the probability distribution for the *sum* ϵ of the kinetic energies of both electrons, $P''(\epsilon)$; and the half-life of two-neutrino $\beta^{\mp}\beta^{\mp}$ decay, $T_{1/2}'$. Thus:

$$C'(\epsilon_1, \epsilon_2, \theta)$$
$$= 1 + \left[\frac{\int_0^{\epsilon^{(0)}-\epsilon_1-\epsilon_2} B'(\epsilon_1, \epsilon_2, k_1, \epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)\, k_1^2(\epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)^2\, dk_1}{\int_0^{\epsilon^{(0)}-\epsilon_1-\epsilon_2} A'(\epsilon_1, \epsilon_2, k_1, \epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)\, k_1^2(\epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)^2\, dk_1}\right]\cos\theta, \qquad \ldots\ldots(68)$$

$$P'(\epsilon_1) = \int_0^{\epsilon^{(0)}}\int_0^{\pi} Q(\mathbf{p}_1, \mathbf{p}_2)\, 2\pi \sin\theta\, d\theta\, d\epsilon_2$$
$$\simeq N(\epsilon_1+1)^2\int_0^{\epsilon^{(0)}-\epsilon_1}(\epsilon_2+1)^2\left[\int_0^{\epsilon^{(0)}-\epsilon_1-\epsilon_2} A'(\epsilon_1, \epsilon_2, k_1, \epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)\right.$$
$$\left.\times\, k_1^2(\epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)^2\, dk_1\right] d\epsilon_2, \qquad \ldots\ldots(69)$$

$$P''(\epsilon) = \int_0^{\epsilon^{(0)}}\int_0^{\epsilon^{(0)}}\int_0^{\pi} Q(\mathbf{p}_1, \mathbf{p}_2)\, \delta(\epsilon^{(0)}-[\epsilon_1+\epsilon_2])\, 2\pi \sin\theta\, d\theta\, d\epsilon_1\, d\epsilon_2$$
$$\simeq N\int_0^{\epsilon^{(0)}}\left[(\epsilon_1+1)^2(\epsilon^{(0)}-\epsilon_1+1)^2\left[\int_0^{\epsilon^{(0)}-\epsilon} A'(\epsilon_1, \epsilon-\epsilon_1, k_1, \epsilon^{(0)}-\epsilon-k_1)\, k_1^2\right.\right.$$
$$\left.\left.\times\,(\epsilon^{(0)}-\epsilon-k_1)^2\, dk_1\right]\right] d\epsilon_1, \qquad \ldots\ldots(70)$$

$$\ln 2/T_{1/2}' = \int_0^{\epsilon^{(0)}} P'(\epsilon_1)\, d\epsilon_1 = \int_0^{\epsilon^{(0)}} P''(\epsilon)\, d\epsilon$$
$$\simeq N\int_0^{\epsilon^{(0)}}(\epsilon_1+1)^2\left[\int_0^{\epsilon^{(0)}-\epsilon_1}(\epsilon_2+1)^2\left[\int_0^{\epsilon^{(0)}-\epsilon_1-\epsilon_2} A'(\epsilon_1, \epsilon_2, k_1, \epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)\, k_1^2\right.\right.$$
$$\left.\left.\times\,(\epsilon^{(0)}-\epsilon_1-\epsilon_2-k_1)^2\, dk_1\right] d\epsilon_2\right] d\epsilon_1. \qquad \ldots\ldots(71)$$

To estimate $T_{1/2}'$, $C'(\epsilon_1, \epsilon_2, \theta)$, $P'(\epsilon_1)$, $P''(\epsilon)$, we use for the nuclear matrix elements occurring in $A'(\epsilon_1, \epsilon_2, k_1, k_2)$ (equation (65)), essentially the same procedure as used in the case of no-neutrino $\beta^{\mp}\beta^{\mp}$ decay (equations (49)); there are here, however, no factors like r_{nm} to be replaced by the nuclear radius. Then

from equations (49*a*) and (49*b*),

$$\left|\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} \right| \Psi_1 \right\rangle\right|^2 \simeq \frac{1}{100} \quad \dagger \qquad \ldots\ldots(72a)$$

and

$$\left|\left\langle \Psi_f^* \left| \sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} \boldsymbol{\sigma}_n \cdot \boldsymbol{\sigma}_m \right| \Psi_1 \right\rangle\right|^2 \simeq \frac{1}{100}. \qquad \ldots\ldots(72b)$$

As already mentioned, it may however happen that the dominant contribution to $|\langle \beta^\mp \beta^\mp; (A, Z \pm 2)|(A, Z)\rangle|^2$, and so to $A'(\epsilon_1, \epsilon_2, k_1, k_2)$ and $B'(\epsilon_1, \epsilon_2, k_1, k_2)$ comes from a *single* low-lying 0^+, *or* 1^+ 'intermediate' nucleus state $\Psi_v \equiv \Psi_w$. Then *either* the 'Fermi' type of nuclear matrix element of equation (72*a*) *or* the 'Gamow–Teller' type of equation (72*b*) is dominant, and, assuming that the numerical estimates of these last equations are still valid, we have from equations (65), (49*c*), (49*d*),

$$\{A'(\epsilon_1, \epsilon_2, k_1, k_2)\}_F \simeq 2^4 |F|^4 (K^2 + L^2 - KL) \left|\left\langle \Psi_w^* \left| \sum_{m=1}^{A} \tau_m^{(\pm)} \right| \Psi_1 \right\rangle\right|^4$$

$$\simeq 2^4 |F|^4 (K^2 + L^2 - KL) \left[\frac{1}{100}\right] \qquad \ldots\ldots(73a)$$

$$\frac{\{B'(\epsilon_1, \epsilon_2, k_1, k_2)\}_F}{\{A'(\epsilon_1, \epsilon_2, k_1, k_2)\}_F} \simeq -\frac{KL}{K^2 + L^2 - KL} \cdot \frac{p_1 p_2}{H_1 H_2}, \qquad \ldots\ldots(73b)$$

or

$$\{A'(\epsilon_1, \epsilon_2, k_1, k_2)\}_{GT} \simeq 2^4 |F|^4 3(K^2 + L^2 + KL) \frac{1}{9} \left|\left\langle \Psi_w^* \left| \sum_{m=1}^{A} \tau_m^{(\pm)} \boldsymbol{\sigma}_m \right| \Psi_1 \right\rangle\right|^4$$

$$\simeq 2^4 |F|^4 3(K^2 + L^2 + KL) \frac{1}{9} \left[\frac{1}{100}\right] \qquad \ldots\ldots(74a)$$

$$\frac{\{B'(\epsilon_1, \epsilon_2, k_1, k_2)\}_{GT}}{\{A'(\epsilon_1, \epsilon_2, k_1, k_2)\}_{GT}} \simeq -\frac{[\frac{2}{3}(K^2 + L^2) + \frac{5}{3} KL]}{K^2 + L^2 + KL} \frac{p_1 p_2}{H_1 H_2}. \qquad \ldots\ldots(74b)$$

We can, in this situation, also consider $\langle W_v - W_1 \rangle$ approximately equal to $W_w - W_1$.

To simplify further, we replace $H_j, |\mathbf{k}_i|$ $(i, j = 1, 2)$ in K, L (equation (66)) by their average values over the probability distribution in momenta of the emitted $e_1^\mp, e_2^\mp, \nu_1, \nu_2$. Thus

$$K \simeq L \simeq \left\langle \frac{2}{(W_w - W_1) + H_1 + |\mathbf{k}_1|} \right\rangle_{av} \simeq \frac{2}{(W_w - W_1) + \frac{1}{2}(\epsilon^{(0)} + 2)}. \qquad \ldots\ldots(75)$$

† It is to be noted that the 'Fermi' type nuclear matrix element in equation (72*a*) (and equation (49*c*)) vanishes if Ψ_1, Ψ_f are characterized by different eigenvalues of the total isobaric spin operator $\mathbf{T}^2$. This follows since

$$\sum_{n,m} \tau_n^{(\pm)} \tau_m^{(\pm)} = 4(T^{(1)} \pm i T^{(2)})^2$$

and so yields non-vanishing matrix elements only between Ψ_1, Ψ_f with the *same* T. Such a situation presumably obtains in $\beta\beta$ decays with low Z where one may suppose that Ψ_1 and Ψ_f are approximately characterized by the smallest T consistent with the corresponding $T^{(3)} = Z - \frac{1}{2}A$. Thus, in $\beta\beta$ decays with low Z, the 'Fermi' type nuclear matrix element is probably considerably smaller than the estimate of $(1/100)^{\frac{1}{2}}$ given in equation (72*a*). On the other hand, for medium and large Z, Ψ_1, Ψ_f are not even approximately eigenstates of $\mathbf{T}^2$ so that the isobaric spin selection rule under discussion ceases to operate and the estimate of equation (72*a*) becomes relevant. Of the $\beta\beta$ decays so far investigated (see table 6, below) it is likely that only ${}^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$ falls into the 'low Z' category.

Equations (75) and (73*a*), (74*a*) yield

$$\{A'(\epsilon_1,\epsilon_2,k_1,k_2)\}_{\mathrm{F}} \simeq \{A'(\epsilon_1,\epsilon_2,k_1,k_2)\}_{\mathrm{GT}}$$

$$\simeq 2^4 |F|^4 \left(\frac{2}{(W_w - W_1) + \frac{1}{2}(\epsilon^{(0)}+2)}\right)^2 \left[\frac{1}{100}\right]. \qquad \ldots\ldots(76)$$

We now use equation (76) for $\{A'(\epsilon_1,\epsilon_2,k_1,k_2)\}_{\mathrm{F,GT}}$ and equations (73*b*), (74*b*), (75) for $\{B'(\epsilon_1,\epsilon_2,k_1,k_2)\}_{\mathrm{F,GT}}$ in equations (68)–(71) to obtain

$$\{C'(\epsilon_1,\epsilon_2,\theta)\}_{\mathrm{F}} \simeq \{C'(\epsilon_1,\epsilon_2,\theta)\}_{\mathrm{GT}} \simeq 1 - \frac{\bar{p}_1 \bar{p}_2}{H_1 H_2}\cos\theta, \qquad \ldots\ldots(77)$$

$$\frac{\{P'(\epsilon_1)\}_{\mathrm{F}}}{N\{A'\}_{\mathrm{F}}} \simeq \frac{\{P'(\epsilon_1)\}_{\mathrm{GT}}}{N\{A'\}_{\mathrm{GT}}} \simeq \frac{(\epsilon_1+1)^2(\epsilon^{(0)}-\epsilon_1)^6}{5040}\{(\epsilon^{(0)}-\epsilon_1)^2 + 8(\epsilon^{(0)}-\epsilon_1) + 28\}, \qquad \ldots\ldots(78)$$

$$\frac{\{P''(\epsilon)\}_{\mathrm{F}}}{N\{A'\}_{\mathrm{F}}} \simeq \frac{\{P''(\epsilon)\}_{\mathrm{GT}}}{N\{A'\}_{\mathrm{GT}}} \simeq \frac{(\epsilon^{(0)}-\epsilon)^5\,\epsilon}{900}\{\epsilon^4 + 10\epsilon^3 + 40\epsilon^2 + 60\epsilon + 30\}, \qquad \ldots\ldots(79)$$

$$\{T_{1/2}{}'\}_{\mathrm{F}} \simeq \{T_{1/2}{}'\}_{\mathrm{GT}} \simeq 6\times 10^{19\pm 2}\left(\frac{(W_w - W_1) + \frac{1}{2}(\epsilon^{(0)}+2)}{10}\right)^2 \left(\frac{60}{Z}\right)^2$$

$$\times \left(1 - \exp\left(\mp\frac{2\pi Z}{137}\right)\right)^2 \left(\frac{8}{\epsilon^{(0)}}\right)^{10} \mathrm{yr}. \qquad \ldots\ldots(80)$$

In equation (80) $\epsilon^{(0)}$ is the (two-neutrino) $\beta^\mp\beta^\mp$ decay energy release defined in §2 and expressed in units of $m_{\mathrm{elec}}c^2$. The coupling constant $|F|$, entering the above expressions, is found from the single $\beta^\mp$ decay data (equation (63*a*) and discussion after equation (55)). The proportionality of the half-lives in equation (80) to $(\epsilon^{(0)})^{-10}$ arises from a very rough approximation for the integral of $P'(\epsilon_1)$ in equation (71). The ± 2 in the exponent is an estimate of the overall error which arises largely from the uncertainty in the nuclear matrix elements (equation (72)).

Comparison of the $1/T_{1/2}$ of equation (54) with the $1/T_{1/2}{}'$ of equation (80) indicates that, for comparable $\epsilon^{(0)}$, Z and nuclear matrix elements, the no-neutrino $\beta^\mp\beta^\mp$ decay rates are faster than the corresponding two-neutrino $\beta^\mp\beta^\mp$ decay rates by a factor 10^5–10^6. From the transition matrix elements for these two types of decay (equations (33) and (60)), and their relation to the corresponding transition probabilities (equations (43) and (67)), it can be seen that this factor is, essentially, the fourth power of the ratio of the energies of the no-neutrino decay's 'virtual' neutrino, and the two-neutrino decay's 'real' neutrino, viz. approximately $[35\ \mathrm{Mev}/1\ \mathrm{Mev}]^4 \simeq 1{\cdot}5\times 10^5$.

The angular correlation function $C'(\epsilon_1,\epsilon_2,\theta)$ and the probability distributions $P'(\epsilon_1)$, $P''(\epsilon)$ (equations (77)–(79)) are all subject to experimental test. It is important to note that while the $T_{1/2}{}'$ are not sensitive to the approximation for K, L in equation (75) (and so are given fairly accurately for a wide range of values of the ratio $(W_w - W_1)/\frac{1}{2}(\epsilon^{(0)}+2)$ by equation (80)), the quantities $C'(\epsilon_1,\epsilon_2,\theta)$, $P'(\epsilon_1)$, $P''(\epsilon)$ are influenced by this approximation, and, unless

$$(W_w - W_1) \gg \tfrac{1}{2}(\epsilon^{(0)}+2),$$

are only very roughly given by equations (77), (78) and (79). In particular, it has been shown (Rosen 1957, Blin-Stoyle and Rosen 1958, unpublished), for the case

$$(W_w - W_1) \ll \tfrac{1}{2}(\epsilon^{(0)} + 2),$$

that if $P'(\epsilon_1)$ and $P''(\epsilon)$ are calculated from equations (69) and (70), using equations (74) and (66), but *not* equation (75), the results are rather different from those in equations (78) and (79). In this calculation the decay ($\epsilon^{(0)} = 3{\cdot}3$ MeV)

$$^{96}_{40}\mathrm{Zr} \rightarrow {}^{96}_{41}\mathrm{Nb} + \mathrm{e}_1^- + \nu_1 \rightarrow {}^{96}_{42}\mathrm{Mo} + \mathrm{e}_1^- + \mathrm{e}_2^- + \nu_1 + \nu_2$$

is considered, and a special nuclear model is used to estimate $W_w - W_1$. The $P''(\epsilon)$ obtained has a broad, skewshaped peak at an ϵ quite close to $\epsilon^{(0)}$, while the $P''(\epsilon)$ of equation (79) has a broad symmetrical peak at an $\epsilon \simeq \epsilon^{(0)}/2$.

3.6. *ββ Decays with Orbital Negatron Capture*

We give, without detailed discussion, estimates for the half-life of a $\beta\beta$ decay with orbital (K-shell) negatron capture. As indicated in §2, such a process may be possible energetically when $\beta^+\beta^+$ decay is energetically impossible and is always possible energetically when $\beta^+\beta^+$ decay is.

With methods similar to those of §§3.4, 3.5, we obtain

(*a*) *No-neutrino* K *capture*, e^+ *emission*

$$T_{1/2} \simeq 5 \times 10^{18\pm2}\left(\frac{60}{Z}\right)^4 \exp\frac{2\pi(Z-60)}{137}\left(\frac{A}{150}\right)^{2/3}\left(\frac{8}{\epsilon^{(0)}}\right)^4 \text{yrs.} \qquad \ldots\ldots(81)$$

(*b*) *Two-neutrino* K *capture*, e^+ *emission*

$$\{T_{1/2}\}_F \simeq \{T_{1/2}\}_{GT} \simeq 3 \times 10^{22\pm2}\left(\frac{W_w - W_1 + \frac{1}{2}\epsilon^{(0)}}{10}\right)^2\left(\frac{60}{Z}\right)^4 \exp\frac{2\pi(Z-60)}{137}\left(\frac{8}{\epsilon^{(0)}}\right)^7 \text{yrs.} \qquad \ldots\ldots(82)$$

(*c*) *No-neutrino* KK *capture*

$$T_{1/2} \simeq 5 \times 10^{23\pm2}\left(\frac{60}{Z}\right)^5\left(\frac{A}{150}\right)^{2/3}\left(\frac{8}{\epsilon^{(0)}}\right) \text{yrs.} \qquad \ldots\ldots(83)$$

(*d*) *Two-neutrino* KK *capture*

$$\{T_{1/2}\}_F \simeq \{T_{1/2}\}_{GT} \simeq 3 \times 10^{25\pm2}\left(\frac{(W_w - W_1) + \frac{1}{2}(\epsilon^{(0)} - 2)}{10}\right)^2\left(\frac{60}{Z}\right)^6\left(\frac{8}{\epsilon^{(0)}}\right)^5 \text{yrs.} \qquad \ldots\ldots(84)$$

In equations (81)–(84), $\epsilon^{(0)}$ is the energy release in units of $m_{\text{elec}}c^2$ (see §2). The proportionality of the half-lives in equations (81), (82), to $(\epsilon^{(0)})^{-4}$ and $(\epsilon^{(0)})^{-7}$ is very rough. The factor ± 2 in all the exponents is again an estimate of the overall error, mainly due, as before, to the uncertainty in the estimates of the nuclear matrix elements in equations (49) and (72).

The e^+ in no-neutrino K capture, e^+ emission is mono-energetic (see §1); also, an inner bremsstrahlung photon is expected to carry off the energy released in no-neutrino KK capture (Winter 1955*b*).

3.7. *Relation of ββ Decay to Inverse β Decay*

To supplement the comments made in §1, we shall discuss the relation of $\beta\beta$ to inverse β decay, considering, for simplicity, the Majorana case ($\nu \equiv \bar{\nu}$).

With the nucleon–lepton interaction described by H_M (equation (16)), we obtain an expression for R, the rate of the inverse β decay reaction already considered in equation (10), viz.

$$\nu^- + n \rightarrow p + e^-, \qquad \ldots\ldots(85)$$

where the $\nu^-(\nu^- = \bar{\nu} \equiv \nu)$ are emitted in the β^- decay of neutron-rich nuclei (pile produced fission fragments)

$$n \rightarrow p + e^- + \nu^-. \qquad \ldots\ldots(86)$$

We now express R as follows. Let $p(\frac{1}{2})$ and $p(-\frac{1}{2})\,[= 1 - p(\frac{1}{2})]$ be the probabilities that the ν^- in equation (86) is emitted with its spin oriented respectively parallel and anti-parallel to its momentum. Further, let $g(\frac{1}{2})$ and $g(-\frac{1}{2})$ be the cross sections in equation (85) for capture by a neutron of a ν^- with spin and momentum oriented respectively parallel and anti-parallel. Then

$$R = (\nu^- \text{ flux}) \times (p(\tfrac{1}{2})g(\tfrac{1}{2}) + p(-\tfrac{1}{2})g(-\tfrac{1}{2})). \qquad \ldots\ldots(87)$$

The expressions $p(\pm\frac{1}{2}), g(\pm\frac{1}{2})$ can be evaluated, and it appears that (Rosen and Primakoff, unpublished, see also Enz 1957, 1958)

$$p(\pm\tfrac{1}{2})g(\pm\tfrac{1}{2}) = \sum_{\mu,\lambda;\,\mu',\lambda'} \mathscr{L}(\mu,\lambda;\,\mu',\lambda')(I_{\mu\lambda} \mp J_{\mu\lambda})(I_{\mu'\lambda'} \mp J_{\mu'\lambda'})^*, \qquad \ldots\ldots(88)$$

where $I_{\mu\lambda}, J_{\mu\lambda}$ are given by equation (34*b*) and $\mathscr{L}(\mu,\lambda;\,\mu',\lambda')$ involves nuclear matrix elements associated with the reactions of equations (85) and (86).

Equations (87) and (88) show that R vanishes when all the $I_{\mu\lambda}, J_{\mu\lambda}$ vanish; however, as seen in §§ 3.4, 3.5, all $I_{\mu\lambda} = J_{\mu\lambda} = 0$ is also the condition for the absence of no-neutrino $\beta\beta$ decay. Conversely, if some of the $I_{\mu\lambda}, J_{\mu\lambda}$ are non-zero, both the inverse β decay of equation (85) and the no-neutrino $\beta\beta$ decay proceed, in general, at a non-vanishing rate. We see, therefore, that there exists, in effect, a one-to-one correspondence between the occurrence or the absence of one process and the occurrence or absence of the other. It is finally worth mentioning that with the 'two-component neutrino' coupling used in §3.5 (H_M of equation (16) with $F_S = F_T = 0; |F_V| \neq 0, |F_A| \neq 0, \epsilon_V = \epsilon_A = 1$), we have

$$p(\tfrac{1}{2}) = 1, p(-\tfrac{1}{2}) = 0;\ g(\tfrac{1}{2}) = 0, g(-\tfrac{1}{2}) \neq 0, \qquad \ldots\ldots(89)$$

so that R vanishes immediately. Equation (89) itself is most easily obtained from equations (16) and (17) by noting that:

ν^- emission

$$\begin{aligned}(1+\gamma_5)\,\mathrm{d}_\nu^\dagger(\mathbf{k},\kappa,s)\mathscr{C}^* u^*(\mathbf{k},\kappa,s;\mathbf{r}) &= \mathrm{d}_\nu^\dagger(\mathbf{k},\kappa,s)\mathscr{C}^*((1-\gamma_5)^* u^*(\mathbf{k},\kappa,s;\mathbf{r})) \\ &= \mathrm{d}_\nu^\dagger(\mathbf{k},\kappa,s)\mathscr{C}^*\left(\left(1+\frac{\boldsymbol{\sigma}.\mathbf{k}}{k}\right)u(\mathbf{k},\kappa,s;\mathbf{r})\right)^* \\ &= \begin{cases} 2\mathrm{d}_\nu^\dagger(\mathbf{k},\kappa,s)\mathscr{C}^* u^*(\mathbf{k},\kappa,s;\mathbf{r}) & \text{if } s = \frac{1}{2} \\ 0 & \text{if } s = -\frac{1}{2}. \end{cases}\end{aligned}$$

ν^- absorption

$$\begin{aligned}(1+\gamma_5)\,\mathrm{d}_\nu(\mathbf{k},\kappa,s)u(\mathbf{k},\kappa,s;\mathbf{r}) &= \mathrm{d}_\nu(\mathbf{k},\kappa,s)\left(1-\frac{\boldsymbol{\sigma}.\mathbf{k}}{k}\right)u(\mathbf{k},\kappa,s;\mathbf{r}) \\ &= \begin{cases} 0 & \text{if } s = \frac{1}{2} \\ 2\mathrm{d}_\nu(\mathbf{k},\kappa,s)u(\mathbf{k},\kappa,s;\mathbf{r}) & \text{if } s = -\frac{1}{2}. \end{cases} \qquad \ldots\ldots(90)\end{aligned}$$

§4. Comparison of Experimental and Theoretical $\beta\beta$ Half-lives

We collect in table 6 the $\beta\beta$ half-lives determined by experiment together with the corresponding theoretical estimates. The experimental values, which apart from three exceptions are lower limits on the half-lives, are those discussed in §2 and the theoretical estimates, both for two-neutrino and no-neutrino $\beta\beta$ decay, are calculated from equations (80), (84) and (54), (83) of §3. (In equations (80) and (84), $[(W_w - W_1) + \frac{1}{2}(\epsilon^{(0)} \pm 2)]/10$ is taken as unity). An examination of the table indicates that most of the evidence favours the two-neutrino process particularly since of the three positive observations of $\beta\beta$ decay two are contradicted by later work ($^{100}_{42}\mathrm{Mo} \rightarrow {}^{100}_{44}\mathrm{Ru}$ and $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$), while the third has an experimental half-life which is in good agreement with the two-neutrino theoretical value ($^{130}_{52}\mathrm{Te} \rightarrow {}^{130}_{54}\mathrm{Xe}$). However, all the half-life estimates are uncertain by a factor of roughly a hundred (§3) and the final decision between two-neutrino and no-neutrino $\beta\beta$ decay will come only when $e_1^{\mp}, e_2^{\mp}$ time-coincident events, uniquely associated with a $\beta^{\mp}\beta^{\mp}$ source, are observed. The criterion will then be whether $P''(\epsilon)$, the probability distribution for the sum ϵ of the kinetic energies of the two emitted electrons, is better approximated, allowing for instrumental resolution, by equation (70) or (79) appropriate to two-neutrino $\beta^{\mp}\beta^{\mp}$ decay, or by $\delta(\epsilon^{(0)} - \epsilon)$ (equation (43)) which applies to no-neutrino $\beta^{\mp}\beta^{\mp}$ decay.

Table 6

Decay process	Energy release, $\epsilon^{(0)}$(MeV)	$T_{\frac{1}{2}}$ (years) Experimental	Theoretical Two-neutrino	Theoretical No-neutrino
$^{130}_{52}\mathrm{Te} \rightarrow {}^{130}_{54}\mathrm{Xe}$	3·0	$= 1{\cdot}4 \times 10^{21}$	$2 \times 10^{21 \pm 2}$	$8 \times 10^{15 \pm 2}$
$^{238}_{92}\mathrm{U} \rightarrow {}^{238}_{94}\mathrm{Pu}$	1·0	$> 6 \times 10^{15}$	$3 \times 10^{25 \pm 2}$	$2 \times 10^{18 \pm 2}$
$^{124}_{50}\mathrm{Sn} \rightarrow {}^{124}_{52}\mathrm{Te}$	2·2	$> 1 \times 10^{16}$ $> 1 \times 10^{17}$ $> 2 \times 10^{17}$ $> 5 \times 10^{16}$ $> 1{\cdot}5 \times 10^{17}$	$4 \times 10^{22 \pm 2}$	$5 \times 10^{16 \pm 2}$
$^{116}_{48}\mathrm{Cd} \rightarrow {}^{116}_{50}\mathrm{Sn}$	2·7	$> 1 \times 10^{17}$ $> 6 \times 10^{16}$	$6 \times 10^{21 \pm 2}$	$2 \times 10^{16 \pm 2}$
$^{100}_{42}\mathrm{Mo} \rightarrow {}^{100}_{44}\mathrm{Ru}$	2·3	$= 2 \times 10^{16}$ $> 3 \times 10^{17}$	$4 \times 10^{22 \pm 2}$	$4 \times 10^{16 \pm 2}$
$^{106}_{48}\mathrm{Cd} \rightarrow {}^{106}_{46}\mathrm{Pd}$	0·9	$> 6 \times 10^{16}$	$2 \times 10^{25 \pm 2}$	$6 \times 10^{20 \pm 2}$
$^{96}_{40}\mathrm{Zr} \rightarrow {}^{96}_{42}\mathrm{Mo}$	3·3	$> 2 \times 10^{16}$ $> 5 \times 10^{17}$	$1 \times 10^{21 \pm 2}$	$6 \times 10^{15 \pm 2}$
$^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$	4·3	$= 2 \times 10^{17}$ $> 2 \times 10^{18}$	$4 \times 10^{20 \pm 2}$†	$3 \times 10^{15 \pm 2}$
$^{150}_{60}\mathrm{Nd} \rightarrow {}^{150}_{62}\mathrm{Sm}$	3·7	$> 4 \times 10^{18}$	$2 \times 10^{20 \pm 2}$	$2 \times 10^{15 \pm 2}$
$^{64}_{30}\mathrm{Zn} \rightarrow {}^{64}_{28}\mathrm{Ni}$ KK capture	1·1	$> 1 \times 10^{16}$	$1 \times 10^{20 \pm 2}$	$3 \times 10^{16 \pm 2}$

† *Note added in proof*: On the basis of a shell model estimate of the relevant nuclear matrix elements, Belyaev and Zaharyev (1958) give $T_{\frac{1}{2}} \simeq 10^{19}$ years for the two-neutrino $\beta^-\beta^-$ decay of $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$.

§ 5. Conclusion

The basic experimental question associated with the $\beta\beta$ decay process is :

(I) Are neutrinos emitted in $\beta\beta$ decay ? The corresponding basic theoretical question is :

(II) Can it be decided, on the basis of the empirically obtained answer to (I), whether the neutrino is, or is not, identical with the anti-neutrino and whether the total lepton charge is, or is not, conserved in the nucleon–lepton interaction ?

To conclude this review we shall give a summary of our answers to these questions.

As indicated in § 2 and in § 4, while it is not yet certain that any $\beta\beta$ decay has been observed, it is nevertheless clear that the existing evidence definitely points towards $\beta\beta$ decay being a two-neutrino rather than a no-neutrino process. Though such a conclusion is supported by the most convincing experiments so far performed, namely those of Inghram and Reynolds (1949, 1950) and Cowan, Harrison, Langer and Reines (1956), the conflict between the results of McCarthy (1955) and of Awschalom (1956) on $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$ is not completely settled, and we believe, especially in view of its large energy release, that this $\beta\beta$ decay should be re-examined. Then, perhaps, it will be possible to give a completely unequivocal answer to question (I).†

As regards question (II), prior to the discovery that parity is not conserved in the nucleon–lepton interaction, one concluded that if $\beta\beta$ decay were of the no-neutrino type, then :

(A) Conservation of total lepton charge was violated ; neutrino $\equiv$ anti-neutrino as the simplest possibility, whereas if two neutrinos were emitted in $\beta\beta$ decay, then,

(B) Total lepton charge was conserved ; neutrino $\not\equiv$ anti-neutrino.

However, as discussed in detail in § 1 and § 3, the situation with a nucleon–lepton interaction without conservation of parity is no longer so clear cut ; in this case, if the $\beta\beta$ decay is no-neutrino the conclusion (A) may still be drawn, but if the $\beta\beta$ process is two-neutrino, conclusion (B) is not ***unique***, conclusion (B′) also being possible, where :

(B′) Total lepton charge was not conserved ; neutrino $\equiv$ anti-neutrino but with a nucleon–lepton interaction with 'two-component neutrino' type coupling. Of course, the nucleon–lepton interaction contemplated in (B) may also involve 'two-component neutrino' type coupling.

The ambiguity between the conclusions (B) and (B′) renders impossible an unique answer to (II) if two-neutrino $\beta\beta$ decay is actually present. The situation may also be viewed as follows. Suppose that further theoretical and experimental developments establish beyond reasonable doubt that the non-conservation of parity in weak interactions involving neutrinos is correctly described by the presence of 'two-component neutrino' type couplings in the corresponding interaction Hamiltonians. Then, consistent with the balance of the above described

† *Note added in proof* : In an elegant experiment, which unfortunately came to the attention of the authors too late for inclusion in § 2, Dobrochotov, Lazarenko and Lukyanov (1956) obtain $T_{\frac{1}{2}} > 1 \times 10^{18}$ years for the $\beta^-\beta^-$ decay in $^{48}_{20}\mathrm{Ca} \rightarrow {}^{48}_{22}\mathrm{Ti}$; this is in agreement with the result of Awschalom (1956) and in disagreement with that of McCarthy (1955). In unpublished work on the same $\beta^-\beta^-$ decay, der Mateosian (1958, private communication) obtains a similar lower limit on $T_{\frac{1}{2}}$. This last pair of experiments further strengthens the evidence in favour of the two-neutrino character of $\beta\beta$ decay.

experimental evidence, only two-neutrino $\beta\beta$ decay is expected, *whether*

(B) total lepton charge is conserved ('Dirac' case, $\nu \not\equiv \bar{\nu}$), *or*

(B′) total lepton charge is not conserved ('Majorana' case, $\nu \equiv \bar{\nu}$).

Decision between (B) and (B′) now necessarily involves the study of neutrino behaviour in reactions *other* than those of $\beta\beta$ decay or inverse β decay. In particular, the Michel ρ value in muon decay, which empirically is approximately 3/4 rather than approximately 0, implies that $\nu \not\equiv \bar{\nu}$ (§ 1) and so excludes (B′). Thus, assuming provisionally the universal applicability of 'two-component neutrino' type coupling and considering the $\beta\beta$ decay, the inverse β decay and the muon decay data together, a verdict may be tentatively reached in favour of a 'Dirac' neutrino, operationally distinguishable from a 'Dirac' anti-neutrino, and with conservation of total lepton charge valid in all neutrino interactions.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research (ARDC). We should like to thank Mrs. Norma Schroeder for help in preparation of the manuscript.

References

Alder, K., Stech, B., and Winther, A., 1957, *Phys. Rev.*, **107**, 728.
Alford, W. P., and Hamilton, D. R., 1957, *Phys. Rev.*, **105**, 673.
Allen, J. S., 1958, *The Neutrino* (New Jersey; Princeton University Press).
Awschalom, M., 1956, *Phys. Rev.*, **101**, 1041.
Bell, J. S., 1955, *Proc. Roy. Soc.* A, **231**, 479.
Belyaev, V. B., and Zaharyev, B. N., 1958, *J. Exp. Theor. Phys.*, **34**, 505.
Berthelot, A., Chaminade, R., Levi, C., and Papineau, L., 1953, *C.R. Acad. Sci., Paris*, **236**, 1769.
Blatt, J., and Weisskopf, V., 1952, *Theoretical Nuclear Physics* (New York: John Wiley).
Bouchiat, C., and Michel, L., 1957, *Phys. Rev.*, **106**, 170.
Burgy, M. T., Krohn, V. E., Novey, T. B., Ringo, G. R., and Telegdi, V. L., 1958, *Phys. Rev.*, **110**, 1214.
Case, K. M., 1957, *Phys. Rev.*, **107**, 307.
Cavanagh, P. E., Coleman, C. F., Ridley, B. W., and Turner, J. F., 1958, *Nuclear Physics*, **5**, 11.
Cowan, C. L., Harrison, F. B., Langer, L. M., and Reines, F., 1956, *Nuovo Cim.*, **3**, 649.
Cowan, C. L., and Reines, F., 1953 a, *Phys. Rev.*, **90**, 492; 1953 b, *Ibid.*, **92**, 830; 1956, *Nature, Lond.*, **178**, 446; 1957, *Phys. Rev.*, **107**, 528.
Crowe, K. M., 1957, *Bull. Amer. Phys. Soc.*, Series II, **2**, 206.
Curtis, R. B., and Lewis, R. R., 1957, *Phys. Rev.*, **107**, 543.
Davis, R., 1955, *Phys. Rev.*, **97**, 766; 1956, *Bull. Amer. Phys. Soc.*, Series II, **1**, 219.
Detoef, J. F., and Moch, R., 1955, *J. Phys. Radium*, **16**, 897.
Dobrohotov, E. I., Lazarenko, B. R., and Lukyanov, S. U., 1956, *Doklady Akad. Nauk. S.S.S.R.*, **110**, 966.
Dudziak, W. F., and Sagane, R., 1957, *Proc. 7th Annual Rochester Conference* (New York: Interscience), VII-25.
Ebel, M. E., and Feldman, G., 1957, *Nuclear Physics*, **4**, 213.
Enz, C. D., 1957, *Nuovo Cim.*, **6**, 250; 1958, *Helv. Phys. Acta*, **31**, 69.
Feenberg, E., 1955, *Shell Theory of the Nucleus* (New Jersey: Princeton University Press), Chapter V.
Feynman, R. P., and Gell-Mann, M., 1958, *Phys. Rev.*, **109**, 193.
Fireman, E. L., 1949, *Phys. Rev.*, **75**, 323.
Fireman, E. L., and Schwarzer, D., 1952, *Phys. Rev.*, **86**, 451.
Fremlin, J. H., and Walter, M. C., 1952, *Proc. Phys. Soc.* A, **65**, 911.

FRIEDMAN, J. I., and TELEGDI, V. L., 1957, *Phys. Rev.*, **105**, 1681.
FULBRIGHT, H., 1952, *Physica 's Grav.*, **18**, 1026 ; 1955, *Beta- and Gamma-ray Spectroscopy*, edit. K. Siegbahn (New York : Interscience), p. 668.
FURRY, W. H., 1939, *Phys. Rev.*, **56**, 1184.
GARWIN, R. L., LEDERMAN, L. M., and WEINRICH, M., 1957, *Phys. Rev.*, **105**, 1415.
GATTO, R., 1958, *Nuclear Physics*, **5**, 530.
GOLDHABER, M., GRODZINS, L., and SUNYAR, A. W., 1958, *Phys. Rev.*, **109**, 1015.
GOOD, M. L., and LAUER, E. J., 1957, *Phys. Rev.*, **105**, 213.
HERMANNSFELDT, W. B., BURMAN, R. L., STÄHELIN, P., ALLEN, J. S., and BRAID, T. H., 1958, *Phys. Rev. Letters*, **1**, 61.
HERMANNSFELDT, W. B., MAXSON, D. R., STÄHELIN, P., and ALLEN, J. A., 1957, *Phys. Rev.*, **107**, 641.
HOUTERMANS, F. G., and THIRRING, W., 1954, *Helv. Phys. Acta*, **27**, 81.
INGHRAM, M. G., and REYNOLDS, J. H., 1949, *Phys. Rev.*, **76**, 1265 ; 1950, *Ibid.*, **78**, 822.
JACKSON, J. D., TREIMAN, S. B., and WYLD, H. W., 1957 a, *Phys. Rev.*, **106**, 517 ; 1957 b, *Nuclear Physics*, **4**, 206.
JOST, R., 1957, *Helv. Phys. Acta*, **30**, 409.
KAHANA, S., and PURSEY, D. L., 1957, *Nuovo Cim.*, **6**, 1469.
KALKSTEIN, M. I., and LIBBY, W. F., 1952, *Phys. Rev.*, **85**, 368.
KINOSHITA, T., and SIRLIN, A., 1957 a, *Phys. Rev.*, **107**, 593 ; 1957 b, *Ibid.*, **107**, 638.
KOHMAN, T. P., 1st March 1954, *U.S. Atomic Energy Commission Report*, NYO-3626.
KONOPINSKI, E. J., 1955, *Los Alamos Report* (LAMS-1949).
KONOPINSKI, E. J., and MAHMOUD, H., 1953, *Phys. Rev.*, **92**, 1045.
KONUMA, M., 1958, *Nuclear Physics*, **5**, 504.
LANDAU, L., 1957, *Nuclear Physics*, **3**, 127.
LARSEN, S., LUBKIN, E., and TAUSNER, M., 1957, *Phys. Rev.*, **107**, 856.
LAWSON, J. S., 1951, *Phys. Rev.*, **81**, 299 (A).
LEE, T. D., 1957, *Proc. 7th Annual Rochester Conference* (New York : Interscience), VII-1.
LEE, T. D., and YANG, C. N., 1956, *Phys. Rev.*, **104**, 258 ; 1957, *Ibid.*, **105**, 1671.
LEE, T. D., OEHME, R., and YANG, C. N., 1957, *Phys. Rev.*, **106**, 340.
LEVENE, C. A., GHIORSO, A., and SEABORG, G. T., 1950, *Phys. Rev.*, **77**, 296.
LEWIS, R., 1957, *Phys. Rev.*, **108**, 904.
LÜDERS, G., 1954, *K. Danske. Vidensk. Selsk., Mat-Fys. Med.*, **28**, No. 5 ; 1957, *Annals of Physics*, **21** ; 1958, *Nuovo Cim.*, **7**, 171.
LÜDERS, G., and ZUMINO, B., 1957, *Phys. Rev.*, **106**, 385.
MARX, G., 1953, *Acta Physiol. Hung.*, **3**, 55 ; 1954, *Z. Naturf.*, **9a**, 105 ; 1957, *Nuovo Cim.*, **6**, 1235.
MAXSON, D. R., ALLEN, J. S., and JENTSCHKE, W., 1955, *Phys. Rev.*, **97**, 109.
MAYER, M. G., 1935, *Phys. Rev.*, **48**, 512.
MAYER, M. G., and TELEGDI, V. L., 1957, *Phys. Rev.*, **107**, 1445.
MCCARTHY, J. A., 1953, *Phys. Rev.*, **90**, 853 ; 1955, *Ibid.*, **97**, 1234.
MCLENNAN, J. A., 1957, *Phys. Rev.*, **106**, 821.
MICHEL, L., 1950, *Proc. Phys. Soc.* A, **63**, 514.
PAULI, W., 1955, *Niels Bohr and the Development of Physics* (New York : McGraw-Hill) ; 1957, *Nuovo Cim.*, **6**, 204.
PEARCE, R. M., and DARBY, E. K., 1952, *Phys. Rev.*, **86**, 368.
PRESTON, M. A., 1957, *Canad. J. Phys.*, **35**, 1017.
PRIMAKOFF, H., 1952, *Phys. Rev.*, **85**, 888 ; 1953, *University of Pittsburgh Conference on Medium Energy Nuclear Physics* (Conference Report).
RACAH, G., 1937, *Nuovo Cim.*, **14**, 322.
RIDLEY, B. W., 1956, *Progr. Nucl. Phys.*, **5**, 188.
ROSEN, S. P., 1957, *Ph.D. Thesis*, University of Oxford.
ROSENSON, L., 1958, *Phys. Rev.*, **109**, 958.
RUSTAD, B. M., and RUBY, S. L., 1955, *Phys. Rev.*, **97**, 991.
SAKURAI, J., 1958, *Nuovo Cim.*, **7**, 649.
SALAM, A., 1957, *Nuovo Cim.*, **5**, 299.

SARGENT, C. P., RINEHART, M., LEDERMAN, L., and ROGERS, K. C., 1955, *Phys. Rev.*, **99**, 885.
SCHWINGER, J., 1951, *Phys. Rev.*, **82**, 914; 1953, *Ibid.*, **91**, 713.
SHERR, R., and MILLER, R. H., 1954, *Phys. Rev.*, **93**, 104.
SLIV, C. A., 1950, *J. Exp. Theor. Phys.*, **20**, 1035.
SUDARSHAN, E. C. G., and MARSHAK, R. E., 1957, *Proc. Padua-Venice International Conference* (Padua; Ciclografia Borghero), V-14; 1958, *Phys. Rev.*, **109**, 1860.
TIOMNO, J., 1950, *Ph.D. Thesis*, Princeton University.
WINTER, R. G., 1955 a, *Phys. Rev.*, **99**, 88; 1955 b, *Ibid.*, **100**, 142.
WU, C. S., AMBLER, E., HAYWARD, R. W., HOPPES, D. D., and HUDSON, R. P., 1957, *Phys. Rev.*, **105**, 1413.
ZELDOVICH, Y. B., 1953, *Doklady Akad. Nauk. S.S.S.R.*, **91**, 1317.

PHYSICAL PROBLEMS OF THERMOELECTRICITY

By A. F. JOFFE AND L. S. STIL'BANS
Academy of Sciences, Leningrad, U.S.S.R.

Translated by M. G. Priestley

CONTENTS

	PAGE
§ 1. Introduction	168
§ 2. The efficiency of thermoelectric generators of electricity, refrigerators and heaters	168
§ 3. The conditions for the optimum figure-of-merit of thermoelements	170
3.1. Mobility and effective mass	172
3.2. Thermal conductivity	174
3.3. The thermoelectric properties of solid solutions	180
§ 4. Establishment of the optimum carrier concentration	185
§ 5. Additional factors relating to the thermal e.m.f.	188
§ 6. The scattering mechanisms of the charge carriers	194
6.1. The scattering of electrons by thermal vibrations of the lattice	194
6.2. The scattering of electrons in solid solutions	199
6.3. The investigation of the mechanism of scattering by ionized impurities	200
References	203

Abstract. The interest in thermoelectric phenomena and in their practical applications led to an investigation of semiconductors with high electric conductivities σ and low thermal conductivities κ. The requirement of the highest possible ratio of the mobility U to κ directed our attention to solid solutions. κ could be decreased to values corresponding to a free path of phonons of a few interatomic distances.

The replacement of positive ions in a compound by analogous ions (for instance Bi by Sb in Bi_2Te_3 or Pb by Sn in PbTe) decreases many times the mobility of electrons, while the mobility of holes remains unaltered. A distortion of the sub-lattice of negative ions leads to a decrease of the mobility of holes.

The influence of impurities on the free path of phonons and of free charge carriers reveals the scattering cross section of defects. Crystal lattices with a proportion of unfilled sites have exceptionally short free paths which, calculated in the usual way, are smaller than the periods of the lattice. Accordingly the mobility increases with rising temperature.

Using degenerate systems of free charge carriers it became possible to separate the influence, on the mechanism of scattering, of the velocity of electrons on one hand, and of the thermal motion of the lattice on the other.

The onset of intrinsic conduction introduces an additional thermal conductivity due to the diffusion of electron–hole pairs. Excitons, with small excitation energies, have an analogous effect.

The high transparency of pure semiconductors to radiation with $h\nu < E_0$ brings into action a heat conduction mechanism, due to the transmission of infra-red radiation, as long as absorption by impurities or by free charges does not stop the radiation.

The dragging of phonons and electrons increases both the thermal conductivity and the thermoelectric power at low temperatures.

The materials selected for thermoelectric devices are investigated in detail.

§ 1. Introduction

The considerable advances made during the last ten years in the electronics of transistors and in the physics of the processes which occur in them are well known.

In connection with this, the overwhelming majority of research into semiconductors in recent years has concerned those processes and materials which are related to the operation of transistors. The main subjects of study have been the elements of the fourth group (and some compounds of the type $A_{III}B_{V}$), and in these substances the main topics of research have been the energy levels, the lifetime, the mobility of carriers, and some galvanomagnetic effects.

One of the basic merits of this work is its strong sense of purpose—which has also been its main drawback—and from the results of this work a very wide class of semiconducting materials has appeared and the nature of many important physical processes which occur in them has been elucidated.

From this point of view the physical research carried out in recent years on the investigation of materials for semiconducting thermoelements in the Institute of Semiconductors of the Academy of Sciences of the U.S.S.R. has been a useful supplement.

The work has to a considerable degree the character of a search and has covered a wide range of materials, thus enabling us to establish a series of laws which characterize the relation between the mechanical, electrical and thermal properties of various substances and their crystallographic and chemical structures. The applications of thermoelectricity demanded a more detailed study of the mechanism of the motion and scattering of electrons and phonons in various substances and under various external conditions.

The aim of the present article is to discuss the formulation of the problems which arise in research in this field and the answers which we have obtained.

§ 2. The Efficiency of Thermoelectric Generators of Electricity, Refrigerators and Heaters

The Seebeck effect, which was discovered in 1821, makes possible the direct partial transformation of thermal into electrical energy without the use of intermediate machines.

It is interesting that among the materials studied by Seebeck were some that could have increased the efficiency of thermoelectric generators to 3%, i.e. to an efficiency equal to that of contemporary steam engines. It is difficult to say how the development of electricity would have progressed had Seebeck's thermoelectric machines been built fifty years before Faraday's development of the induction motor. But this did not happen, and now thermoelectric generators are much less efficient than magnetic generators, and only the extreme simplicity of thermoelements ensures for them the possibility of practical application in low-energy work.

As well as their cost and technology, the size and weight of thermoelements are of great importance. Nevertheless, their efficiency is most important, and is determined both by the temperatures T_1 and T_2 between which the thermal battery operates and by the physical properties of the materials of the thermoelements.

As can be shown by an elementary energy calculation, the properties of the materials in both junctions of the thermocouple may be included in one quantity Z, which determines the usefulness of the thermal battery

$$Z = \frac{(\alpha_1 - \alpha_2)^2}{[(\kappa_1 \rho_1)^{1/2} + (\kappa_2 \rho_2)^{1/2}]^2} \deg^{-1} \qquad \ldots\ldots(\text{I})$$

where α_1 and α_2 are the thermal e.m.f. coefficients of the junctions, κ_1 and κ_2 are their specific thermal conductivities, and ρ_1 and ρ_2 are the specific resistances.

The maximum efficiency of the battery can be written as a function of Z, T_1 and T_2 as

$$\eta = \frac{T_1 - T_2}{T_1} \frac{[1 + \frac{1}{2}(T_1 + T_2) Z]^{1/2} - 1}{[1 + \frac{1}{2}(T_1 + T_2) Z]^{1/2} + T_2/T_1}. \qquad \ldots\ldots(\text{II})$$

Consideration of thermoelectric cooling leads to the requirement that the ratio between the amount of heat removed and the expenditure of electrical energy should be a maximum. This ratio ϵ can be called the coefficient of performance of a refrigerator.

A calculation shows that ϵ is given by

$$\epsilon = \frac{T_2}{T_1 - T_2} \frac{[1 + \frac{1}{2}(T_1 + T_2) Z]^{1/2} - T_1/T_2}{[1 + \frac{1}{2}(T_1 + T_2) Z]^{1/2} + 1} \qquad \ldots\ldots(\text{III})$$

where Z has the same value as in formula (I).

Finally, the problem of heating leads to the following expression for the efficiency (in this case the ratio of the heat obtained to the electrical energy used to produce it)

$$K = \epsilon + 1. \qquad \ldots\ldots(\text{IV})$$

Thus the efficiency of all thermoelectric machines is completely determined by the quantity Z. Below we shall consider what conditions lead to a maximum value of Z.

We note that in formulae (II), (III) and (IV) the quantity Z always appears as a multiplying factor of $(T_1 + T_2)/2 = T_{av}$, i.e. the average absolute temperature of the thermoelement. Since Z has the dimensions $\deg^{-1}$, the product ZT_{av} is a dimensionless ratio independent of the system of measurement. Only this product appears in the expressions (II), (III), (IV), which characterize the successes which have been obtained in the increasing of the efficiency of a thermal battery, using semiconductors with a given Z over a given temperature interval.

If we write equation (II) in the form

$$\eta = \left(1 - \frac{T_2}{T_1}\right) \frac{(1 + ZT_{av})^{1/2} - 1}{(1 + ZT_{av})^{1/2} + T_2/T_1} \qquad \ldots\ldots(\text{IIa})$$

we see that η is a function of the two dimensionless quantities ZT_{av} and T_2/T_1.

Formula (III) can be written in the same form:

$$\epsilon = \frac{1}{T_1/T_2 - 1} \frac{(1 + ZT_{av})^{1/2} - T_1/T_2}{(1 + ZT_{av})^{1/2} + 1}. \qquad \ldots\ldots(\text{IIIa})$$

Therefore in the estimation of the value of a given thermoelectric apparatus it is better to use the quantity ZT_{av}. For the thermoelements constructed at present by us $ZT_{av} \simeq 1$. In the English literature the property of the materials known as the 'figure of merit' is determined by $Z^{1/2}$.

§ 3. The Conditions for the Optimum Figure-of-merit of Thermoelements

For simplicity we shall in future assume $|\alpha_1| = |\alpha_2|$, $\kappa_1 = \kappa_2$, and $\rho_1 = \rho_2 = 1/\sigma$, where σ is the specific electrical conductivity, and then for each of the junctions

$$Z = \alpha^2 \sigma / \kappa. \qquad \ldots\ldots(1a)$$

The quantity Z represents the value of a given substance as a material for thermoelements.

We shall calculate the conditions for which Z reaches its maximum value. The values of the quantities which appear in the expression for Z, the thermal e.m.f. coefficient α, the specific thermal conductivity κ, and the electrical conductivity σ, are not independent, and they are all functions of the concentration of free charge carriers or (what is the same thing) of the position of the level of the chemical potential μ or of the reduced chemical potential $\mu^* = \mu/\boldsymbol{k}T$, where $\boldsymbol{k}$ is the Boltzmann constant.

$$\alpha = \frac{\boldsymbol{k}}{\boldsymbol{e}} \left(\frac{r+1}{r+2} \frac{F_{r+1}(\mu^*)}{F_r(\mu^*)} - \mu^* \right), \qquad \ldots\ldots(2)$$

$$n = \frac{4\pi (2m\boldsymbol{k}T)^{3/2}}{\boldsymbol{h}^3} F_{1/2}(\mu^*), \qquad \ldots\ldots(3)$$

$$\sigma = \frac{16\pi \boldsymbol{e}^2 m l_0(T) (\boldsymbol{k}T)^{r+1}}{3\boldsymbol{h}^3} F_r(\mu^*), \qquad \ldots\ldots(4)$$

$$\kappa_{el} = \left(\frac{\boldsymbol{k}}{\boldsymbol{e}} \right)^2 AT\sigma, \qquad \ldots\ldots(5)$$

where

$$A = \begin{cases} \frac{1}{3}\pi^2 \text{ for the degenerate case,} \\ r+2 \text{ for non-degenerate charges.} \end{cases}$$

m = effective mass

Here we introduce the notation :

(i) l is the mean free path of an electron as a function of the energy and the temperature ; in the general case this dependence is usually expressed as

$$l = l_0(T)\, \epsilon^r$$

where the exponent r of ϵ is determined by the scattering mechanism for electrons (for an atomic lattice $r = 0$, for an ionic lattice below the Debye temperature calculation gives $r = \frac{1}{2}$, above the Debye temperature $r = 1$, for scattering by impurity ions $r = 2$, etc.).

(ii) $F_{1/2}(\mu^*)$, $F_1(\mu^*)$, etc. are the Fermi integrals

$$F_r(\mu^*) = \int_0^\infty x^r \phi dx \qquad \ldots\ldots(6)$$

where $x = \epsilon/kT$ is the reduced kinetic energy and ϕ is the Fermi distribution function

$$\phi = \frac{1}{\exp(x-\mu^*)+1}.$$

The Fermi integrals have been tabulated and the problem of the determination of Z as a function of μ^* and the other parameters which determine σ and κ (T, m, r, etc.) can be solved graphically.

However, such an approach would not give an answer to the following interesting question: what conditions must be satisfied by the parameters of the materials of the junctions for Z to be a maximum?

To answer this question we must find an analytical solution, and to do this we must make a series of assumptions which will simplify the solution of the problem.† It can be shown that Z, considered as a function of μ^* (or n), always has a maximum.

(i) We assume that this maximum occurs for carrier concentrations which are such that classical statistics are still applicable, and in this case the Fermi distribution function may be replaced by the Maxwell distribution

$$\phi \simeq \exp[(\mu-\epsilon)/kT] \qquad \ldots\ldots(7)$$

and the expressions for the thermal e.m.f., the electrical conductivity, the carrier concentration, and the constant A in the Wiedemann–Franz law are considerably simplified, μ^*, α, σ and κ can be expressed as elementary functions of n:

$$\mu^* = \ln \frac{h^3 n}{2(2\pi m kT)^{3/2}}$$

$$\alpha = \pm\frac{k}{e}(r+2-\mu^*) = \pm\frac{k}{e}\left[r+2+\ln\frac{2(2\pi m kT)^{3/2}}{h^3 n}\right]$$

$$\sigma = enU,$$

$$A = r+2.$$

(ii) We further assume that for carrier concentrations which correspond to the maximum value of Z, the electronic part of the thermal conductivity κ_{el} is small compared with the thermal conductivity of the lattice κ_{ph}. Then

$$Z = \frac{\alpha^2\sigma}{\kappa_{el}+\kappa_{ph}} \simeq \frac{\alpha^2\sigma}{\kappa_{ph}}. \qquad \ldots\ldots(8)$$

(iii) We further assume that the thermal conductivity of the crystal lattice κ_{ph} and the mobility U of the carriers are independent of their concentration. With these simplifying assumptions we may easily determine the carrier concentration n_0 which corresponds to Z_{max}. Actually we have, in this case,

$$\frac{\partial Z}{\partial n} = \frac{\partial(\alpha^2\sigma)}{\partial n}$$

$$\alpha^2\sigma = \frac{k^2}{e^2}\left[r+2+\ln\frac{2(2\pi m kT)^{3/2}}{h^3 n}\right]^2 nU, \qquad \ldots\ldots(9)$$

† We shall explain below how far these assumptions are justified.

and the condition $\partial(\alpha^2\sigma)/\partial n = 0$ gives the optimum carrier concentration and the corresponding electrical conductivity

$$n_0 = \frac{2(2\pi m \boldsymbol{k} T)^{3/2}}{\boldsymbol{h}^3} e^r, \quad \sigma = eUn_0 \qquad \ldots\ldots(10)$$

and the optimum thermal e.m.f.

$$\alpha = 2\frac{\boldsymbol{k}}{\boldsymbol{e}} = 172\,\mu\text{v deg}^{-1}. \qquad \ldots\ldots(11)$$

Hence the expression for $\alpha^2\sigma$ becomes

$$\alpha^2\sigma \simeq 1{\cdot}2 \times 10^{-7} U \left(\frac{m}{m_0}\frac{T}{T_0}\right)^{3/2} e^r, \qquad \ldots\ldots(12)$$

and correspondingly the maximum value of Z is

$$Z_{\max} \simeq 1{\cdot}2 \times 10^{-7} \frac{U}{\kappa_{ph}} \left(\frac{m}{m_0}\frac{T}{T_0}\right)^{3/2} e^r. \qquad \ldots\ldots(13)$$

As can be seen from (10), the optimum carrier concentration depends on several factors: the effective mass, the temperature and the scattering mechanism. By (11) the optimum thermal e.m.f. is the same for all materials under all conditions and equals 172 μv deg^{-1}. Therefore it is expedient to use the thermal e.m.f. as a guide to the optimum carrier concentration, especially since its measurement is so much easier than that of n. By (13) the efficiency Z of a thermocouple is proportional to the ratio of the carrier mobility to the thermal conductivity of the crystal lattice; Z reaches its maximum value for a definite carrier concentration which corresponds to (10). Thus the above analysis shows that the choice of materials for the junctions of thermoelements reduces to the solution of the following problems: (i) The finding of substances which have the highest ratio of carrier mobility to the thermal conductivity of the crystal lattice; (ii) the production of carrier concentration corresponding to (10) in these materials.

We examine separately the methods of solving each of these problems.

The methods of trying to find materials which have a maximum ratio of U/κ_{ph} are still to a large extent obscure. This is explained by the insufficiency of our theoretical and experimental knowledge of the dependence of the electrical and thermal properties of crystals on their chemical and crystallographic structure. We can change the properties of a substance in a desirable manner to a small extent by the action of external influences (mechanical, irradiation, etc.). At the present time the usual method is by chemical purification or the introduction of impurities. The near future will undoubtedly give us much important information for the solution of these problems. Here we shall try to explain in qualitative form what we know at present in this field.

3.1. *Mobility and Effective Mass*

The mobility of carriers of electricity is limited by their scattering by thermal vibrations and by lattice defects. We shall not now discuss the effects of defects, since we can reduce their number, and consequently their influence on the mobility,

to a certain extent, by the choice of the method of preparation of the materials† (by the purification necessary for the synthesis of materials, by growing single crystals, by annealing, etc.). Consequently we shall consider only scattering by the thermal vibrations of the lattice.

Here it must be borne in mind that the scattering is not caused by the thermal vibrations of the atoms themselves, but by the disturbances of the periodic potential which they cause. The stronger the disturbance of the periodic potential, i.e. roughly speaking, the larger its amplitude, the larger will be the local fields which arise from the thermal vibrations. Therefore the mobility is basically determined by the following factors :

(i) The Debye temperature θ : the lower θ the greater will be the intensity of thermal vibrations at a given temperature, and consequently the smaller will be the mobility.

(ii) The degree to which the compound is ionic : the mobility of carriers is usually low in ionic compounds where the potential variations are most marked ; on the other hand, the carrier mobility is considerably higher in crystals with binding of a covalent character (where the alternation of positive and negative charged ions is absent). Comparison of a series of isomorphous crystals, such as, for example, PbS, PbSe, PbTe, clearly illustrates how the mobility increases with the transition from ionic to covalent binding.

(iii) In the series of compounds quoted above the increase in mobility is actually only partially explained by a reduction in the degree of ionic character of the binding—the same phenomenon is observed in the series diamond–silicon–germanium–grey tin, in which there is no question of an ionic bond.

The second reason for the increase in mobility on moving to heavier elements is that the polarizability of the atoms, or ions, in one sub-group of the periodic table, but in different periods, increases as their atomic number increases (i.e. as their electronic envelope grows and the number of electrons contained in it increases). Heavy atoms are polarized to a considerable extent not only by the influence of external fields, but also by the action of the periodic potential in the crystal, and thus the potential variations are smoothed out ; the heavier the atoms which form the covalent crystal, the less definite and the more smeared out is the covalent binding. It can be stated that the binding becomes less well defined as the atoms become heavier.

We shall now discuss how Z depends on the effective mass.

In the expression (13) for Z the effective mass m^* appears, first, directly, on account of the dependence of the thermal e.m.f. on m^*, and, second, indirectly, because the mobility of the carriers also depends on their effective mass. In accordance with the theory of crystals with ionic binding $U \sim m^{*-3/2}$ and Z is therefore independent of m^*. In the more important case of covalent binding $U \sim m^{*-5/2}$ and, by (13), Z is inversely proportional to the effective mass.

It should be noted that for a number of reasons it is not advisable to attach much significance to this prediction of the theory (the relative dependence of U on m^*). In particular the present theory does not take into account the relationship

† This does not refer to those defects which unavoidably appear when we introduce into the material a certain quantity of donors or acceptors which are necessary to give the optimum carrier concentration ; we shall return to this point later.

between the deformation potential and the effective mass which certainly exists. This dependence can be very strong.

3.2. *Thermal Conductivity*

The thermal conductivity of the materials from which a thermoelement is made plays a decisive role. In semiconductors it consists of two terms: of κ_{ph} the heat transfer by thermal vibrations of the molecules or phonons, and of κ_{el} the heat transfer by free charges.

The basic conceptions of κ_{ph} were expressed in 1913 by Born and Debye and were developed into a quantum theory by Peierls in 1929.

The experimental values were obtained by Eucken. According to Debye the phonon part of the thermal conductivity can be written in the form

$$\kappa_{ph} = \tfrac{1}{3} c v \bar{\lambda}$$

where c is the thermal capacity of 1 cm³ of the material, v is the group velocity of propagation of thermal vibrations, which is the velocity of sound, $\bar{\lambda}$ is the average mean free path for phonons. For amorphous solids $\bar{\lambda}$ is of the order of one or two intermolecular distances. In crystals $\bar{\lambda}$ reaches a value of ten or a hundred times the lattice constant.

The quantity $\bar{\lambda}$ is limited by phonon scattering, due to fluctuations of the thermal motion or crystal defects.

Peierls showed that the basic scattering process is the interaction between phonons, in which two phonons create a third or a phonon splits into two. Above the Debye temperature the scattering probability is proportional to the absolute temperature T. As the thermal capacity and the velocity of sound do not depend strongly on the temperature in solid bodies, we may conclude, in good agreement with experiment, that the phonon thermal conductivity of materials above their Debye temperature is inversely proportional to the temperature T.

Impurity and defect scattering is proportional to the concentration and depends little on the temperature.

At low temperatures $\bar{\lambda}$ is limited by phonon scattering at grain boundaries of the material. Here the dragging of phonons by electrons can play an important role. Phonons which have long wavelengths are very weakly scattered, have a long mean free path, and in the presence of a temperature gradient they acquire a preferential direction of motion along the gradient.

Recently the processes which occur at very low temperatures have been studied both theoretically and experimentally, and extremely interesting results have been obtained. However, since present-day thermoelements are not used at these temperatures, we shall not give an account of these results here. On the other hand the temperature range most interesting for studies of thermoelements, between 100 and 1000°K, has been studied very little. We quote the most important results which have been obtained over the last few years. On the basis of existing conceptions it could have been expected that in every material the phonon thermal conductivity decreases with a rise in temperature, and that for different substances it is larger at a given temperature for those which have the higher Debye temperature and for those in which the anharmonic character of the thermal motion, which creates the interaction between phonons, is expressed more weakly.

Since the Debye temperature θ is determined by the frequency of natural vibrations of the atoms or molecules of the body, it could be expected that θ, and with it κ_{ph}, would be larger for a lower atomic or molecular weight of the substance or for stronger binding, i.e. for a higher modulus of elasticity.

As can easily be seen from the data shown in figure 1, these predictions are justified: the phonon thermal conductivity decreases with increase in atomic weight A and decreases with the transition from weakly anharmonic covalent crystals to ionic compounds. For the same atomic weight κ_{ph} decreases for the transition from the elements of the fourth group of Mendeleev's periodic table to the III–V compounds, then to the II–VI and finally to the I–VII compounds.

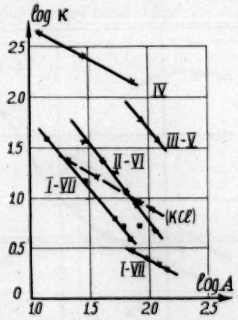


Figure 1.

The effect of anharmonic terms shows up most clearly if the values of κ_{ph} are compared at the Debye temperature, at which the thermal motion is of equal intensity in all substances.

Impurities lower the thermal conductivity of semiconductors by reducing the mean free path of phonons. For concentrations which are not too large their influence can be expressed by the equation

$$\frac{\kappa_0}{\kappa} = 1 + \frac{N}{N_0}\frac{\lambda_0}{a}S$$

where κ_0 is the thermal conductivity of the pure substance, N is the number of impurity atoms, and N_0 is the number of atoms in the crystal, λ_0 is the mean free path of phonons in the pure substance, and S is the scattering cross section which one atom presents to a phonon, expressed in units of the area of a cell of the crystallographic lattice a^2; a is the period of the lattice.

Experiments showed that for substitutional impurities S is close to unity, but that interstitial impurities lead to a value of S approaching 10.

For N/N_0 approaching unity, κ_{ph} falls to a few atomic distances and from the point of view of its thermal conductivity a solid solution of two crystals approaches an amorphous material.

Thus the formation of a solid solution enables us to reduce $\bar{\lambda}$ by a large factor and correspondingly to reduce κ_{ph}. It is necessary, however, to take into account the changes in other properties, in particular in the charge mobility.

There are crystals with a stoichiometric composition in which there are lattice sites not occupied by atoms. For example, in the crystals Ga_2Se_3, Ga_2Te_3, In_2Te_3, which crystallize in a lattice of the zinc blende type, one third of all the sites corresponding to metallic atoms are statistically unoccupied. In these crystals $\bar{\lambda}$ has an extremely low value and may even be less than the lattice constant a. In such cases the whole conception of phonons becomes meaningless and other processes for transferring heat between the molecules of a solid body must be considered.

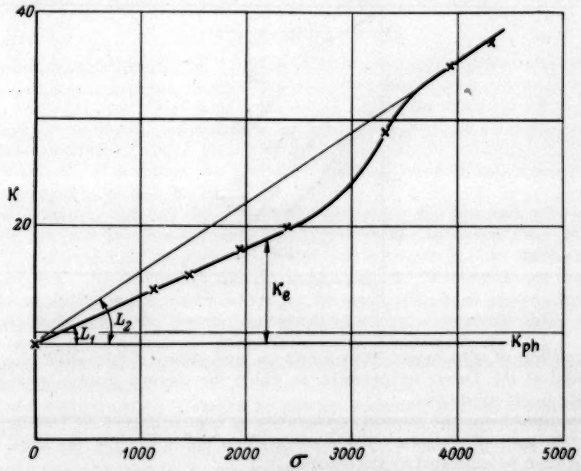


Figure 2.

In semiconductors, the electronic part of the thermal conductivity κ_{el}, which is related to the electrical conductivity σ, acquires considerable importance.

In semiconductors with current carriers of one sign this relation follows from the Wiedemann–Franz law

$$\kappa_{el} = \frac{k^2}{e^2} AT\sigma$$

where $A = \pi^2/3$ for full degeneracy, and $A = r+2$ for non-degenerate current carriers; r denotes the power to which ϵ is raised in the expression for the mean free path l: $l \sim \epsilon^r$.

For the semiconductors which were used to check this dependence r was equal to 0 and hence A equals 2 in the non-degenerate case.

Figure 2 shows the relation between κ and σ for impurity conductivity of compounds of the type Bi_2Te_3. For $\sigma < 2{\cdot}5 \times 10^3$ ohm^{-1} cm^{-1} the straight line has a slope near to $2Tk^2/e^2$ and κ can be expressed as $\kappa = \kappa_{ph} + 2T\sigma k^2/e^2$ where κ_{ph} has the constant value $9{\cdot}6 \times 10^{-3}$ watts cm^{-1} deg^{-1}; for $\sigma > 3{\cdot}5 \times 10^3$ ohm^{-1} cm^{-1} $\kappa = \kappa_{ph} + 3T\sigma k^2/e^2$, which shows the presence of degeneracy. The appearance of free charges of a different sign and in particular of the intrinsic conductivity of the semiconductor introduces a new consideration. The common diffusion of electrons and holes does not create, as for charges of one sign, an opposing diffusion of the electric field. If the concentration of charges of both signs decreases in the direction of the temperature gradient, then they recombine during the passage from higher to lower temperatures, thereby liberating an energy which corresponds to the width of the forbidden energy band ΔE_0.

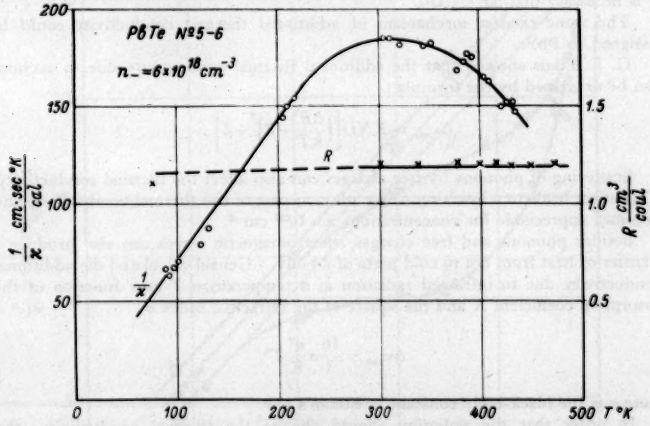


Figure 3.

As long ago as 1940 B. I. Davydov and M. L. Shmushkevich calculated the addition to the thermal conductivity which is due to this mechanism of energy transfer

$$\Delta\kappa = L\frac{\sigma_1\sigma_2}{\sigma_1+\sigma_2}\left(\frac{\Delta E_0}{2kT}+r+2\right)^2.$$

Here L is the ratio κ/σ in the Wiedemann–Franz law, σ_1 and σ_2 are the electrical conductivities for holes and electrons, considered separately, and r is defined as before. The formula of Davydov and Shmushkevich was neglected and was again deduced by Price in 1955. It was confirmed experimentally on materials of the type Bi_2Te_3, where as the intrinsic conductivity appeared an additional conductivity$_h$ κ_p was discovered which agrees well with that calculated.

However the investigation of the temperature dependence for other semiconductors had shown the appearance of an additional thermal conductivity

long before the onset of intrinsic conductivity. Thus, in PbTe the additional conductivity is already appreciable at 200°K and grows exponentially with temperature, whilst measurements of the Hall R effect and the thermoelectric power show that the intrinsic conductivity is appreciable only above 500°K (figure 3).

In this case we may ascribe the presence of an extra $\Delta\kappa$ to the presence of states of excitation in the energy spectrum of PbTe with energies ΔE of the order of 0·1 ev. Since a possibility of diffusion must be attributed to these states, excitons must be assigned to them which differ from excitons of the Mott type, for which ΔE is close to ΔE_0. An elementary calculation shows that for the excitons, whose concentration is determined by the factor $\exp(-\Delta E/kT)$, to be able to play an appreciable part in the thermal conductivity compared with electrons and holes, whose concentration is governed by the factor $\exp(-\Delta E_0/2kT)$, it is necessary that $\Delta E < \frac{1}{2}\Delta E_0$.

The same exciton mechanism of additional thermal conductivity could be assigned to PbSe.

G. E. Pikus showed that the additional thermal conductivity due to excitons can be expressed by the formula :

$$\Delta\kappa_{\text{exc}} = KND\left[\left(\frac{\Delta E'}{kT}+2\right)^2+2\right].$$

Scattering of phonons by free charges can also affect the thermal conductivity. This effect leads to a lowering of the phonon part of the thermal conductivity and becomes appreciable for concentrations $n > 10^{20}$ cm^{-3}.

Besides phonons and free charges, electromagnetic waves can also produce a transfer of heat from hot to cold parts of a body. Genzel calculated the additional conductivity due to infra-red radiation at a temperature T as a function of the absorption coefficient K and the square of the refractive index n :

$$\Delta\kappa_{\text{rad}} = \frac{16}{3}a\frac{n^2}{K}T^3.$$

Here a is the black-body constant in Stefan's law.

In order that the radiation should change the thermal conductivity, the absorption coefficient K must be extremely small. In semiconductors with a forbidden energy band of width ΔE_0 radiation of frequency $\nu = E_0/n$ can be absorbed only by impurities, lattice defects, or free charges. Therefore infra-red radiation may have a considerable effect on the thermal conductivity even for temperatures which are not very high, if the semiconductors are chemically pure and have a small concentration of free charges.

E. D. Devyatkova and I. A. Smirnov, while studying the thermal conductivity of Te as a function of temperature, obtained the results given in figure 4 for its dependence on charge concentration : 10^{15} cm^{-3} (curve I), 10^{17} cm^{-3} (curve II), and 10^{18} cm^{-3} (curve III). Using the dependence of the absorption coefficient on the carrier concentration established by Fan, Moiges showed that these curves were in full agreement with the formula of Genzel.

Thus the problem of the thermal conductivity of semiconductors turns out to be related not only to the thermal properties of the substance, but also to the

detailed picture of its electrical and optical properties, and also to the nature of the chemical binding and the structure of the crystals.

Whilst in several semiconductors with a small atomic weight and covalent binding the phonon part of the thermal conductivity may already, at room temperature, exceed the thermal conductivity of many metals, it falls by two orders of magnitude for substances with a large atomic weight and for a considerable anharmonic content in the thermal vibrations. In solid solutions and defect crystal lattices the phonon conductivity becomes many times smaller.

The general tendency towards a reduction in thermal conductivity with increase in temperature and towards an increase in conductivity for a decrease in temperature may be distorted as a result of the action of free charges or infra-red radiation.

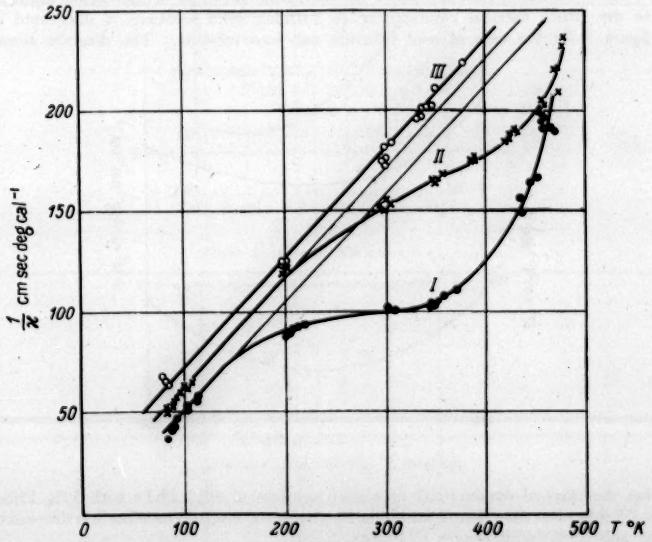


Figure 4.

Very marked effects are observed in chemically pure and physically perfect semiconductors, where the thermal conductivity can be increased many times by the removal of defects at low temperatures and of transparency to infra-red radiation at high temperatures. Heat transfer by free charges also acquires considerable significance.

In any substance the ratio of the carrier mobility to the thermal conductivity of the crystal lattice could be increased by the introduction of neutral impurities which are effective scatterers of elastic waves and which will therefore considerably

reduce the thermal conductivity, but which are not effective electron scatterers (the wavelength of electrons in non-degenerate semiconductors at $T = 300°K$ is at least an order of magnitude greater than the lattice constant, and therefore lattice vibrations which are not propagated over a considerable volume will not be effective for electron scattering).

In particular this can be achieved by the formation of solid solutions of isomorphous compounds. This method of increasing the efficiency of semiconductor thermoelements (suggested by A. V. Joffe) has proved extremely important and deserves separate consideration.

3.3. *The Thermoelectric Properties of Solid Solutions*

The method mentioned above for increasing the ratio of the carrier mobility to the lattice thermal conductivity by forming solid solutions is illustrated in figure 5 for the case of lead telluride and lead selenide. The diagram shows

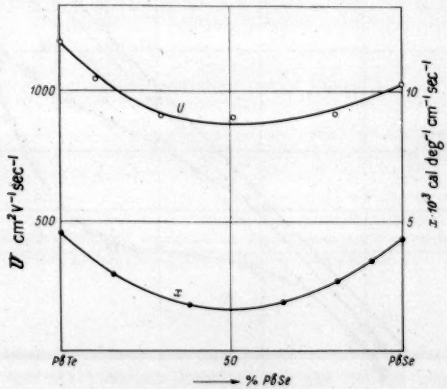


Figure 5. (Symbol like x here is κ).

that the thermal conductivity of a solid solution of 50% PbTe with 50% PbSe is 2·5 times less than that of lead telluride, while the electron mobility has decreased by only 25% (with respect to PbTe).

Figure 6 shows that this increase in the ratio U/κ in a solid solution also holds for the alloys of Bi_2Te_3 with Bi_2Se_3; here the electron mobility is at first unchanged and then increases sharply as the bismuth selenide content is increased, while the thermal conductivity is reduced by a factor of 1·5 for the composition containing 50% Bi_2Te_3.†

† The increase in mobility in this case is due to the fact that the effective mass of the electrons in bismuth selenide is approximately one-third of that in bismuth telluride (see figure 7).

Figure 7 also illustrates clearly the fact that an increase in mobility due to a decrease in the effective mass does not always have a favourable effect on the efficiency of the material $\alpha^2\sigma$. Thus, in this case the composition with a maximum effective mass, and not maximum mobility, corresponds to the maximum value of $\alpha^2\sigma$.

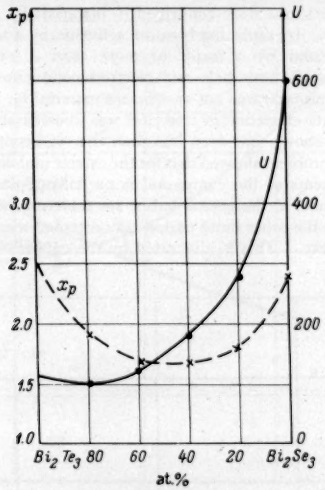


Figure 6. (Symbol like x here is $\varkappa$.)

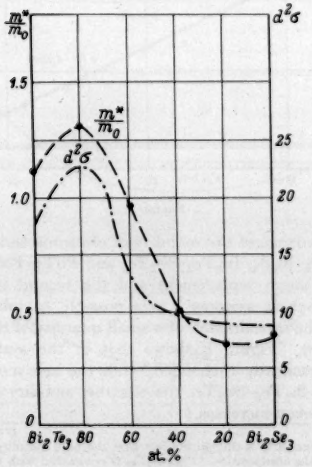


Figure 7.

The same considerations were confirmed in materials for the positive junction of the thermoelement; by replacing bismuth telluride by a solid solution Bi_2Te_3–Sb_2Te_3 Z was increased by a factor of more than 2. However the system Bi_2Te_3–Sb_2Te_3 (in which both hole and electron conduction can be created by suitable choice of impurities) was not an efficient material for the negative junction of thermoelements—its efficiency in this case was considerably less than the two solid solutions given above, and even less than that of bismuth or lead telluride. Further study of this problem showed that for the carrier mobility it is not immaterial which of the components of the compound is partially replaced by its analogue; when some of the anions in the crystal lattice are replaced the mobility of holes is reduced more, but on the other hand partial cation replacement has a greater effect on the electron mobility. This is illustrated by the curves in figures 8, 9 and 10,

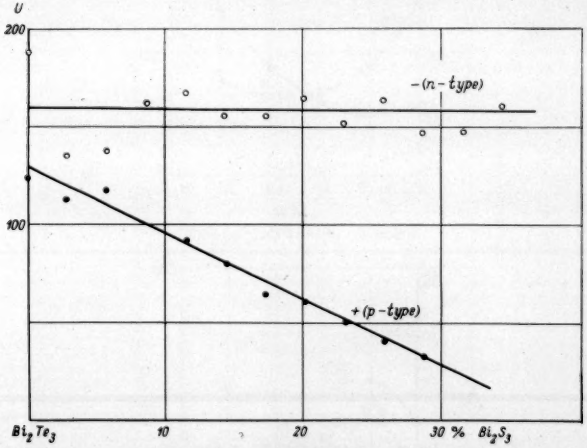


Figure 8.

which show the dependence of the mobility of electrons and holes on composition for the systems Bi_2Te_3–Bi_2S_3, Bi_2Te_3–Sb_2Te_3 and PbTe–PbSe (the first and third cases involve partial anion replacement and the second involves partial cation replacement). In all these systems it was possible to produce the electron or hole conductivity by the introduction of a small quantity of the requisite impurities (see the next section). Figure 8 shows that in the system Bi_2Te_3–Bi_2S_3 the electron mobility is practically unchanged, while the hole mobility falls by a factor of 6. In the system Bi_2Te_3–Sb_2Te_3 the electron mobility falls by a factor of 2, but the hole mobility even increases.†

† Here the increase in mobility is due to the fact that the hole mobility in Sb_2Te_3 is greater than in Bi_2Te_3 (a maximum for the composition 33% Sb_2Te_3 is connected with the formation of an ordered solid solution at this composition).

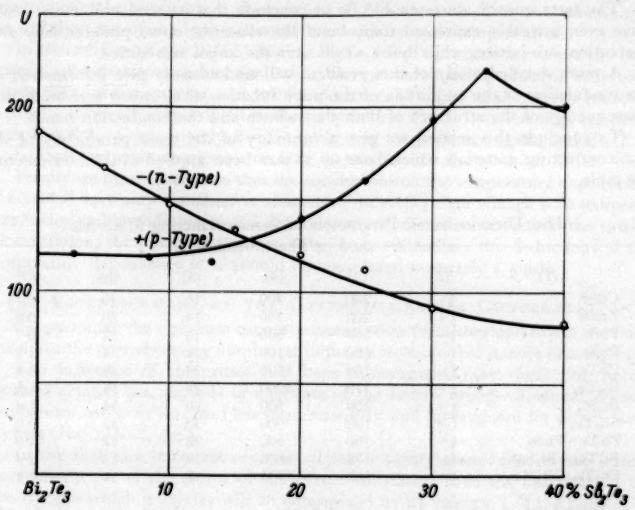


Figure 9.

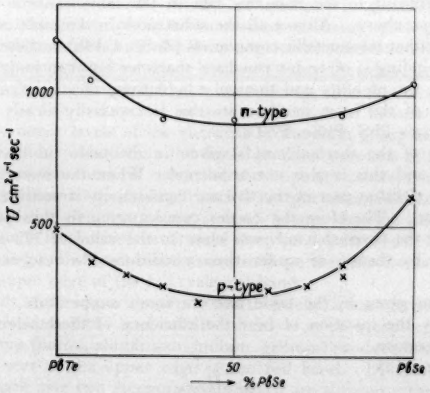


Figure 10.

The facts quoted above enable us to conclude that in semiconductors which have even a weakly expressed ionic bond the electrons move preferentially over the cation sub-lattice, while holes move over the anion sub-lattice.

A more detailed study of this problem will undoubtedly give information on the modulation of the amplitude of the wave function of electrons and holes, and consequently of the structure of both the valence and the conduction bands.

To conclude this section we give a summary of the main parameters of the semiconducting materials which have up to now been applied in thermoelements, see table.

The Thermoelectric Parameters of Semiconducting Materials which are used in Thermoelements

(1)	(2)	(3)	(4)	(5)	(6)	(7)
CdSb	(−)	200	400	16	40	0·4
ZnSb	(+)	200	350	14	20	0·8
PbS	(−)	160	600	16	22	0·7
PbSe	(−)	160	1100	28	24	1·1
PbTe	(−)	160	1500	39	28	1·4
Bi_2Te_3	(−)	170	1000	29	18	1·6
Bi_2Se_3	(−)	200	500	20	14	1·4
Sb_2Te_3	(+)	100	3000	30	32	0·9
PbTe − PbSe	(−)	160	900	23	12	1·9
$Bi_2Te_3 - Bi_2Se_3$	(−)	170	1000	29	12	2·4
$Sb_2Te_3 - Bi_2Te_3$	(+)	160	1500	39	14	2·8

(1) Composition of material; (2) sign of conduction; (3) α (μv deg^{-1}); (4) σ (ohm^{-1} cm^{-1}); (5) $\alpha^2\sigma \times 10^3$ (v^2 g^{-2} ohm^{-1} cm^{-1}); (6) $\kappa \times 10^3$ (w cm^{-1} g^{-1}); (7) $Z \times 10^3$ (deg^{-1}). The values of α, σ and κ are given at room temperature.

It is not difficult to see that the data in the table confirm the theoretical predictions given above. Almost all the substances in the table consist of heavy atoms and are either intermetallic compounds (ZnSb, CdSb) or they are compounds in which the binding is of an intermediate character between ionic and covalent; here the larger the mobility and thermal conductivity of a substance, the weaker the expression of the ionic bond in it—this is especially clearly visible in the homologous series PbS, PbSe, PbTe.

The values of the thermal e.m.fs. given in the table lie between 150 and 200 μv deg^{-1}, and this is also not accidental. When the materials are weakly degenerate the electron part of the thermal conductivity is considerably less than the phonon part. Therefore the carrier concentration in these substances was chosen so that the thermal e.m.f. was close to the value of 172 μv deg^{-1} which was suggested by theory, as under these conditions $\alpha^2\sigma$ reaches its maximum value.

All the data given in the table refer to room temperature. We shall now consider briefly the question of how the efficiency of thermoelements Z varies with the temperature.

According to (13)

$$Z \sim \frac{U}{\kappa_{ph}} T^{3/2} \qquad \ldots\ldots(14)$$

consequently the temperature dependence of Z is determined by the variation of the mobility and of the lattice thermal conductivity with temperature.

In the solid solutions Bi_2Te_3–Bi_2Se_3, Bi_2Te_3–Sb_2Te_3, which are as yet the best materials for thermoelements, the lattice thermal conductivity changes little over a wide range of temperature; at low temperatures ($T < 250°K$) the carrier mobility is usually inversely proportional to the temperature $U \sim T^{-1}$, and at high temperatures $U \sim T^{-2}$. Consequently $Z \sim T^{1/2}$ in the low temperature region, and $Z \sim T^{-1/2}$ in the high temperature region, reaching its maximum value at about 250°K.

Finally we must emphasize that the conclusions on the temperature dependence of U and Z have only a qualitative character: in reality their change with temperature varies within wide limits and depends on a whole series of factors (the carrier concentration, the purity of the materials, etc.). Therefore the deductions of the temperature dependence of Z should be considered as merely a guide.

§ 4. Establishment of the Optimum Carrier Concentration

To guarantee the optimum carrier concentration for a thermoelement we must introduce the corresponding number of impurity atoms (either donors or acceptors).

The influence of impurities has been investigated most fully for purely covalent crystals (i.e. crystals of elements of the fourth group—diamond, silicon, germanium and grey tin), and has been relatively well investigated for purely ionic crystals (NaCl, KCl, etc.).

In the first case (covalent crystals), if the impurity atom 'fits' in the lattice, i.e. replaces one of the atoms of the parent substance, then the character of the free charges which it creates will be determined by its valency. If the number of valence electrons of the impurity atom is greater than four it will be a donor; four electrons are needed for the covalent bonds with the other atoms, while the remainder are loosely bound and need only a small energy increment to bring them into the conduction band. Thus, for example, in germanium, atoms of elements of the fifth group (phosphorus, arsenic, antimony) each give one donor level in the forbidden band near to the lower edge of the conduction band. Elements of the sixth group contribute two donor levels one below the other, i.e. the ionization of the second electron needs considerably more energy than that of the first (which corresponds to the first and second ionization potentials of these atoms *in vacuo*). Even the upper donor levels of the elements of the sixth group are considerably deeper in the forbidden band than the donor levels of elements of the fifth group (this is a result of the chemical nature of the impurity atoms); the increase in electronegativity observed on passing from left to right across Mendeleev's periodic table means that the elements of the seventh group, chlorine and iodine, only produce one donor level which is situated deep in the forbidden band. In this case the remaining ionization potentials are so large that the corresponding levels are below the upper edge of the full (valence) band.

If the valency of the impurity atom is less than four, it has not enough electrons to satisfy its valence bonds: such an atom behaves as an acceptor, thus elements of the third group (boron, aluminium, indium, gallium) produce one acceptor level which is very near to the upper edge of the full band. Divalent metals (zinc, gold, copper) each give two acceptor levels which are situated rather higher, with one above the other (in this case there is a correlation between the position of the

acceptor levels and the electron affinity of the corresponding element). Transition metals (iron, cobalt, nickel, manganese) each give three acceptor levels.

Atoms of several light metals (for example, lithium and copper), which have a small radius, can occupy interstitial positions in the lattices of elements of the fourth group. In this case they easily give up their valence electrons, i.e. they act as donors.

When impurities are introduced into ionic compounds two cases must be distinguished :

1. If the impurity atoms replace a certain number of atoms of one of the components of the compound, then their action is determined (as it is in covalent crystals) by their valency. Thus, for example, if some of the atoms of a divalent metal are replaced by those of a trivalent metal the corresponding number of donor levels will appear in the crystal ; if a divalent metal is partially replaced by univalent atoms the same number of acceptor levels will be formed.

Non-metallic impurities behave in an analogous way : if one of the components of the compound is a metalloid of the sixth group of the periodic system and some of its atoms are replaced by those of an atom of the seventh group, then donor levels will be formed ; when a hexavalent element is replaced by a pentavalent, acceptor levels will appear.

Because of this it should not be thought that the impurity atom itself is necessarily the donor or acceptor—in some cases this is in fact true, but in others the mechanism is more complex. Thus, for example, if in a crystal of any oxide we succeeded in replacing part of the oxygen by chlorine, and this produced an electronic conductivity, it would not follow that the chlorine had given up one of its electrons into the conduction band. What actually happens in such a case is that the atoms of the metal which surround the chlorine have one extra weakly bound electron (i.e. the oxygen ' takes two electrons from its neighbours, and chlorine takes one ') ; this electron state gives a donor level near the lower edge of the conduction band.

2. If the impurity is introduced as an addition to the stoichiometric composition of the ionic compound, then metallic impurity atoms will act as donors irrespective of their valency, and non-metals will act as acceptors ; even if there are no external impurities in the crystal but the ratio of the components differs from the stoichiometric composition an analogous effect is observed—an excess of the metal produces electronic conductivity, and an excess of the non-metal gives rise to hole conductivity. Here, too, it should not be thought that the impurity atom itself always acts as the donor or acceptor : this only happens in those rare cases when it occupies an interstitial position, or in defect crystals (i.e. those in which some of the sites are unoccupied when the stoichiometric composition is satisfied). In several cases the mechanism differs from these.

Thus, for example, after heating an alkali halide crystal in the vapour of alkali metals, an excess of metal and an electronic conductivity was observed in it ; however, measurements of the density of such crystals showed that the departure from stoichiometry was due to the fact that some of the sites of the metalloid sub-lattice were unoccupied ; the atoms of the metal surrounding such an empty site have one extra electron, which ' replaces ' the missing negative ion. This electron state is localized about the vacant metalloid site and gives a donor level in the forbidden band.

In exactly the same way heating of an alkali halide crystal in a halogen vapour creates vacant sites in the metallic sub-lattice which manifest themselves as acceptor levels.

In conclusion we must point out that there are a number of cases where the behaviour of the impurities in covalent and ionic crystals does not agree with the qualitative considerations given above, and needs further and more detailed study.

As we have already noted, the majority of materials for thermoelements occupy an intermediate position between crystals which have purely ionic bonds and those which have purely covalent bonds ; therefore the behaviour of impurities in them is more complex than in the two limiting cases discussed above. We shall consider this question for the compounds which have been most fully investigated : lead telluride and bismuth telluride.

1. As regards the departure from stoichiometric composition, lead telluride behaves in the same way as an ionic compound : an excess of lead produces electronic conductivity, an excess of tellurium hole conductivity. However, the solubility of both components is here very limited ($\sim 0{\cdot}1\%$), and therefore the departure from stoichiometric composition is not sufficient to create the optimum carrier concentration (corresponding to (10)) in the material. The replacement of tellurium or lead atoms by atoms with a different number of valence electrons is considerably more effective.

In this case, too, lead telluride behaves analogously to an ionic compound : replacement of any of the components by an element with a greater number of valence electrons, i.e. of tellurium by one of the halogens, or of lead by a metal of the fifth group (bismuth, antimony or niobium) produces electronic conductivity, in which each impurity atom gives one electron in the conduction band ; when lead is replaced by a metal which lies in the left-hand half of Mendeleev's periodic table holes are produced.

The influence of various impurities on the mobility of electrons and holes was investigated, and here the same effects were observed as were established for solid solutions of isomorphous compounds (see the previous section) : metal impurity atoms which are localized in the lead sub-lattice lower the mobility of electrons, while metalloid impurities lower the hole mobility. Therefore when lead sulphide is being used as a material for the negative junction of a thermo-element, it is advisable to create the optimum carrier concentration by partial replacement of the tellurium by a halogen, rather than by replacing lead with an element from the fifth group, since in the first case the electron mobility will be higher than in the second.

The influence on the concentration and mobility of carriers of impurities of other elements, which do not enter into the chemical composition (and which are not introduced to replace one of the components) of lead telluride, has as yet been very little investigated ; it is relatively well established that here bismuth and antimony act as donors (they dissolve in PbTe to quite a considerable extent), and that the elements of the third group act as acceptors.

By combining what has been said above on the influence of impurities on the electrical properties of lead telluride we can conclude that PbTe behaves from this point of view analogously to a purely ionic compound : metal impurities are always

localized in the cation sub-lattice, and metalloid in the anion sub-lattice; an excess of lead produces electronic conductivity, an excess of tellurium hole conductivity.

2. Bismuth telluride, on the other hand, is in this respect more like a covalent crystal. In this compound the acceptor or donor action of the impurity atom is completely determined by its valency: elements of the seventh and sixth groups (including an excess of tellurium beyond the stoichiometric proportions) produce donor levels, while elements with a smaller number of electrons (including excess bismuth) are acceptors.† If here we compare the influence of an impurity on the concentration of carriers and on their mobility, then it is possible to draw more or less convincing conclusions about the location of the impurity atoms in the crystal lattice. Thus, for example, excess bismuth (i) gives one acceptor level per atom, (ii) considerably decreases the hole mobility, (iii) affects the electron mobility very little. All these three points indicate that the excess bismuth atoms partially replace tellurium in the anion sub-lattice. On the other hand excess tellurium (i) gives one electron per atom, (ii) considerably lowers the electron mobility and (iii) has little effect on the hole mobility, i.e. all these point to localization in the cation sub-lattice. Excess iodine gives 1·5 electrons per atom and has approximately the same influence on the mobility of both electrons and holes; it can therefore be assumed that the iodine atoms are distributed more or less equally over both sub-lattices and give two electrons when they replace bismuth, one when they replace tellurium. Evidently lead partially replaces bismuth, since its atoms are acceptors and sharply reduce the electron mobility. Silver is a donor but affects equally the mobility of both electrons and holes, and has an anomalously large rate of diffusion; all this is in favour of the conclusion that the silver atoms occupy interstitial positions in the bismuth telluride lattice.

This is basically all that can be said at present about the choice of materials for thermoelements and the production in them of the optimum carrier concentration. As we have already noted, and as is clear to the reader from what has been said above, these questions have been as yet very little studied. We hope that in the near future the thermoelectric properties of a much wider range of semiconductors will be investigated, and that the influence of simple and compound impurities on these materials will also be studied. If this work is in fact carried out, it could lead to more effective materials for thermoelements and also to a deeper understanding of the basic laws which determine the electrical and thermal properties of a solid body.

§5. Additional Factors relating to the Thermal E.M.F.

In this section we consider the physical processes which produce an electrical field in a medium where there is a temperature gradient.

In a non-degenerate conductor with a concentration n of free charge carriers which have a mobility U and a diffusion coefficient D, a temperature gradient creates an electric field necessary to balance the flow of charge. The equilibrium

† Several metals of the first group (for example silver) which evidently occupy interstitial positions and act as donors are exceptions to this.

condition can be written in the form

$$EUn = D\frac{dn}{dx} + n\frac{dD}{dx}.$$

The quantities U and D are related by Einstein's equation

$$\frac{U}{D} = \frac{e}{kT}$$

and depend on the scattering mechanism, which is usually expressed in terms of the dependence of the mean free path of the charges l on their average kinetic energy :

$$l \sim \epsilon^r, \quad U \sim \frac{l}{v} \sim \epsilon^{r-1/2}, \quad D \sim TU \sim \epsilon^{r+1/2}.$$

It is also necessary to take into account the variation with temperature of the contact potential ϕ. This is determined by the change in the chemical potential μ.

Thus the thermal e.m.f. $\alpha = \partial E/\partial T$ can be written as

$$\alpha = \frac{k}{e}\left(r + 2 - \frac{\mu}{kT}\right). \qquad \ldots\ldots(15)$$

As can be seen, α depends on only two properties of the semiconductor : on the scattering mechanism and on the chemical potential.

In metals and semi-metals the thermal e.m.f. is determined by the quantity $\alpha \sim \partial g(\epsilon)/\partial \epsilon$, i.e. by the density of states near the Fermi surface.

Since in metals the phonon part of the thermal conductivity is small compared with the electron part, and the latter is determined by the Wiedemann–Franz law, the efficiency of a metallic thermoelement depends only on the value of α. Unfortunately, the question of the conditions on which the energy level density depends, and of the methods of changing it, have not yet been studied. For the opposite case, when $\kappa_{ph} \gg \kappa_{el}$ the optimum efficiency is determined by the maximum of $\alpha^2\sigma$ which is obtained for $\mu = 0$, when the semiconductor is in a state intermediate between degeneracy and non-degeneracy.

Other things being equal, α increases with m^*, but this does not mean that the efficiency increases.

As can be seen from the equation for α, the temperature dependence of the free charge concentration does not appear in the expression (15). On the other hand, the dependence of the diffusion coefficient (and the mobility) on temperature, or on the average kinetic energy ϵ of non-degenerate carriers, plays a considerable role. If the scattering mechanism is such that $r > -\frac{1}{2}$, then D increases with temperature, if $r = -\frac{1}{2}$, D is independent of temperature.

The charge scattering mechanism is an important factor which affects the thermal e.m.f. The experiments described later showed, however, that the replacement of a scattering mechanism with a small r by one for which the value of r equal to 2 was expected, did not have the desired effect. This shows that a more detailed study of the mechanisms of scattering is necessary.

However, before going on to discuss the question of scattering mechanisms we shall consider two further phenomena which affect the value of α.

There is an additional source of thermal e.m.fs. which is known as the phonon drag effect; at low temperatures this effect can give values of the thermal e.m.f. which are hundreds of times greater than those considered above. We assumed that, in the presence of a temperature gradient, only the electron distribution function changes from point to point, but that in each small volume the distribution of velocities in space remains approximately isotropic. Actually, because of the electron–phonon interaction, the latter condition is not obeyed, at least at low temperatures.

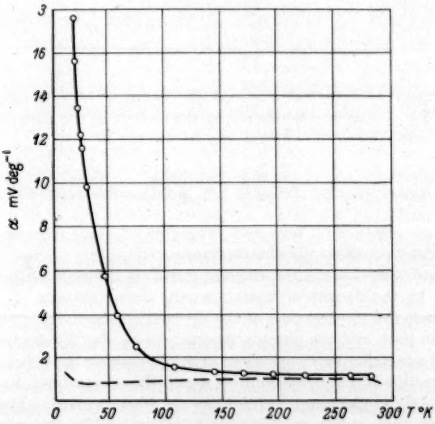


Figure 11.

This phenomenon can be explained qualitatively in the following way:

If there is a temperature gradient in a solid body the number of phonons moving from the hot end to the cold end will be greater than the number moving in the opposite direction. As a result of collisions the phonons will drag electrons along in their own direction of motion, and at the cold end of the specimen a negative charge will appear (similarly a positive charge will appear at the hot end) until the potential difference produced balances the effect of charge flow. This potential difference will be an additional thermal e.m.f. If there is an electrical current (but no thermal gradient) in the body the electrons will carry phonons along with them, which cause an additional heat transfer (the Peltier effect).

The history of the discovery of this effect is instructive. The possibility of the existence of the effect in metals was first established by L. E. Gurevich in 1945, who also showed that the effect was possible in semiconductors. In 1951 Pikus considered this phenomenon qualitatively in semiconductors and deduced

a formula for the additional thermal e.m.f. (due to electrons being carried along by phonons)

$$\alpha_{ph} = \frac{1}{3}\frac{k}{e} m \frac{v_{ph}^2}{k} \frac{1}{T} \frac{\tau_{ph}}{\tau_{el}} \qquad \ldots\ldots(16)$$

where v_{ph} is the velocity of sound, τ_{ph} is the phonon relaxation time, and τ_{el} is the electron relaxation time.

The quantity mv_{ph}^2/k has the dimensions of a temperature, namely that at which the thermal velocity of electrons equals that of phonons. It is not difficult to see that this quantity is of the order of 1°K ; $k/e \simeq 90\ \mu v\ deg^{-1}$ and consequently

$$\alpha_{ph} \simeq \frac{30}{T} \frac{\tau_{ph}}{\tau_{el}}; \qquad \ldots\ldots(17)$$

the value of τ_{ph} can be obtained from the thermal conductivity of the crystal lattice if we use the formula for κ_{ph}

$$\kappa_{ph} = \tfrac{1}{3} C l_{ph} v_{ph} = \tfrac{1}{3} C v_{ph}^2 \tau_{ph}.$$

The value of τ_{el} can be estimated if we know the electron mobility $U = \tau_{ph} e/m$. Having carried out this analysis and having convinced himself that τ_{ph} is not usually more than an order of magnitude different from τ_{el}, Pikus reached the conclusion that α_{ph} cannot play an important role.

However, in 1953, Frederikse and Geballe discovered, while measuring the thermal e.m.f. of germanium, that at low temperatures it increases to values which are many times greater than the predictions of the usual theory, as developed above. Figure 11 shows these data, the broken line showing the value of α according to formula (15).

Hull and Geballe in 1955 found that in p-silicon the thermal e.m.f. reaches even greater values, of the order of 50 mv deg^{-1}.

The theoretical explanation of these anomalously large values of the thermal e.m.fs. was given in 1953 by Herring and Frederikse on the basis of the dragging of electrons by phonons and formula (16), which was first deduced, but incorrectly integrated, by Pikus.

The fact is that (16) must include the average relaxation time not for all the phonons but only for those which can interact with electrons. It can be shown that the laws of conservation of momentum and energy for collisions between electrons and phonons lead to the condition that the phonon wave vector q must satisfy the rule $q \leqslant 2k$ where k is the electron wave vector. Consequently, the wavelength of such phonons $\lambda = 1/q$ must be of the same order as, or greater than, that of the electrons.

At room temperature the wavelength of thermal electrons is of order of magnitude 10^{-7} cm, and at lower temperatures it grows correspondingly. Hence it follows that (16) must contain the relaxation time for phonons whose wavelength is much greater than the lattice constant. Herring showed that the relaxation time τ_{ph}' for such phonons is several orders of magnitude greater than the average relaxation time τ_{ph} obtained from the formula for the thermal conductivity, and consequently α_{ph} is greater than the value Pikus expected by the same factor.

Herring also obtained the temperature dependence of τ_{ph}'. At low temperatures $\tau_{ph}' \sim T^{-3}$ and this gradually changes to $\tau_{ph}' \sim T^{-1}$ at high temperatures.

This dependence agrees well with the experimental data. In those cases where it is necessary to take into account electron and phonon scattering by impurities or at the boundaries of the crystal, the value of α_{ph} drops sharply. Under these conditions formula (16) is not applicable and the expression for α_{ph} becomes considerably more complex. We shall not discuss this case here.

We shall show how formula (16) can be deduced from simple and obvious considerations. If the temperature of the whole body is constant, the elastic waves and the phonons move in random directions; the spatial distribution of phonon velocities is isotropic. If there is a temperature gradient then, as we have already noted, the number of phonons moving from the hot end to the cold is greater than that moving in the opposite direction; we may say that the disordered motion of the phonons becomes a drift in a thermal field (analogously to the way in which the random motion of electrons becomes a drift in the presence of a potential gradient).

We can express the heat flux Q carried by the phonons in terms of this drift velocity v_d and the energy per unit volume $E = CT$ (where C is the thermal capacity) $Q = Ev_d = CTv_d$; on the other hand,

$$Q = \kappa \frac{dT}{dx} = \tfrac{1}{3} C v_{ph} l_{ph} \frac{dT}{dx},$$

and by comparing these two equations we get

$$v_d = \frac{1}{3} \frac{v_{ph} l_{ph}}{T} \frac{dT}{dx}. \qquad \ldots\ldots(18)$$

If the electrons interact only with phonons (and are not scattered by any kind of lattice defects), they will also acquire a drift velocity v_d and will collect at the cold end of the conductor until the electric field which they form produces a drift velocity UE_{ph} which is equal and opposite to v_d. Consequently, in equilibrium,

$$UE_{ph} = v_d,$$

$$E_{ph} = \frac{1}{3} \frac{v_{ph} l_{ph}}{TU} \frac{dT}{dx},$$

$$\alpha_{ph} = \frac{1}{3} \frac{v_{ph} l_{ph}}{TU}. \qquad \ldots\ldots(19)$$

Substituting in (19) $\quad l_{ph} = \tau_{ph} v_{ph}$ and $U = \dfrac{e}{m} \tau_{el}$,

we get

$$\alpha_{ph} = \frac{1}{3} \frac{k}{e} \frac{m v_{ph}^2}{kT} \frac{\tau_{ph}}{\tau_{el}}, \qquad \ldots\ldots(20)$$

which is in full agreement with (16).

It is understood that all the arguments given above refer to the long wavelength phonons and the corresponding value of τ_{ph} should be substituted in (20).

In the consideration of thermoelectric effects connected with the interaction of electrons and phonons there are two possible approaches: the α-approach (i.e. the deductions made above) and the Π-approach. We shall show that the same equation (16) can also be obtained by the second method. To do this we assume

that there is no temperature gradient, but that there is a potential gradient. Here the electrons and phonons change places: the electrons move in the direction of the potential gradient and carry along with them the long wavelength phonons which interact with them, thus creating a flow of heat which is proportional to the electric current (the supplementary Peltier effect). By calculating this heat flow and dividing it by the current we obtain the supplementary Peltier coefficient Π_{ph}.

$$Q = \sum_{q=0}^{q=2k} \epsilon_{(q)} v_{(q)} = \overline{v^2} \sum_{q=0}^{q=2k} hq = \overline{v^2} P_q \qquad \ldots\ldots(21)$$

where ϵ is the phonon energy $\epsilon = h\omega = hqv$, and $P_q = \Sigma hq$ is the total momentum of the phonons which interact with the electrons. In the calculation of $\overline{v^2}$ we must average over all possible directions, therefore $\overline{v^2} = \frac{1}{3} v_{ph}^2$ and we now have to calculate P_q. To do this we examine in detail the mechanism of the interaction between electrons and long wavelength phonons, and between long wavelength and short wavelength phonons. Electrons are accelerated by the electric field and acquire in unit time the momentum $\Delta P_e = eE_n$, and in collisions they transfer this momentum to long wavelength phonons, which then lose it in collisions with short-wave phonons.†

Consequently, in stationary conditions,

$$\frac{dP_q}{dt} = eEn - \frac{P_q}{\tau_{ph}} = 0,$$

whence $P_q = \tau_{ph} eEn$, and since $E = j/eUn$, then $P_q = j\tau_{ph}/U$ and according to (21)

$$Q = \frac{1}{3} \frac{v_{ph}^2 \tau_{ph}}{U} j$$

whence

$$\Pi_{ph} = \frac{Q}{j} = \frac{1}{3} \frac{v_{ph}^2 \tau_{ph}}{U} = \frac{1}{3} \frac{v_{ph} l_{ph}}{U}$$

and

$$\alpha = \frac{\Pi_{ph}}{T} = \frac{1}{3} \frac{v_{ph} l_{ph}}{UT},$$

which is in full agreement with (19).

A special source of thermal e.m.fs. occurs when the local equilibrium of the charge carriers is upset. Up to now we have assumed that the carrier concentration at any point corresponds to equilibrium at the given temperature. Actually this is not always true: if the temperature gradient is so large that the drop in temperature in a diffusion length becomes comparable with the absolute temperature, then the carrier concentration is 'transferred' from the hot regions to the cold ones; this effect is analogous to the volume photovoltaic effect (the Dember effect) and can lead to values of the thermal e.m.f. which are many times different from the predictions of the usual theory.

This effect is especially prominent if large temperature gradients are created in a barrier layer. In this case the excess majority carriers are opposed by the field of the barrier layer, while the minority carriers are accelerated by it, and enormous thermal e.m.fs. are produced which are completely analogous to the

† The short wavelength phonons in their turn give up the momentum to the crystal lattice.

e.m.fs. in photo-diodes. These effects were recently predicted theoretically and then discovered experimentally by the Czech physicist Tauc. Unfortunately, they have as yet been very little studied, and what has been said above is almost all that is known about them.

§6. The Scattering Mechanisms of the Charge Carriers

We have already noted above that the scattering mechanism of carriers has a decisive significance in those properties of a semiconductor which determine its efficiency as a material for thermoelements. The scattering mechanism determines the value of the mobility and its temperature dependence, the magnitude of the thermal e.m.f., and the average energy in the current flow which enters into the Peltier effect. Below we shall consider separately the scattering mechanisms for electrons by the thermal vibrations of the lattice, by neutral and by ionized impurities and by lattice defects.

6.1. *The Scattering of Electrons by Thermal Vibrations of the Lattice*

For the majority of semiconductors which are used as materials for the junctions of thermoelements the temperature dependence of the mobility is given by the expression

$$U \sim T^{-3/2} \qquad \ldots\ldots(22)$$

in spite of the theoretical predictions of

$$U \sim T^{-5/2} \qquad \ldots\ldots(23)$$

for atomic semiconductors and

$$U \sim T^{-1/2}$$

for ionic (above the Debye temperature).

For scattering of a non-degenerate electron gas by lattice thermal vibrations, there are two factors which govern the temperature dependence of the electron mobilities: (i) directly, as a consequence of the fact that the intensity of thermal vibrations increases with temperature; (ii) indirectly, because the mean free path of an electron l depends on its energy, and the average energy in non-degenerate semiconductors is proportional to the temperature. In accordance with this, l may in general always be written as the product of two functions

$$l = \Phi(T)\phi(\epsilon). \qquad \ldots\ldots(24)$$

Theory always gives $\phi(\epsilon)$ in the form $\phi(\epsilon) \sim \epsilon^r$ where the index r depends on the carrier scattering mechanism.

If in the determination of the relaxation time of the electron gas, $\tau = l/v$, we use perturbation theory and, as is usually done, in the calculation of the perturbing potential we take only the first term in the expansion,† then it follows from the theory that only one of the normal vibrations of the lattice takes part in each interaction of the electron with the thermal vibrations, and that the electron can

† i.e. we look for it in the form:

$$U(r) = v(r) - v_0(r) = -(u - \operatorname{grad} v) \qquad (26a)$$

where $U(r)$ is the perturbing potential, $v_0(r)$ is the potential in the undeformed lattice, v is the potential in the lattice with the atoms displaced and u is the displacement of the atom.

receive or give up only one phonon. For this case the laws of conservation of energy and momentum are written as

$$\mathbf{hk}' = \mathbf{hk} \pm \mathbf{hq}$$

$$\frac{\hbar^2 k'^2}{2m} = \frac{\hbar^2 k^2}{2m^*} \pm \hbar\omega \qquad \ldots\ldots(25)$$

where $\mathbf{k}$ and $\mathbf{k}'$ are the wave vectors of the electron before and after the collision, $\mathbf{q}$ is the wave vector of the phonon, ω is the frequency of the normal vibrations, and m^* is the effective mass of the electron. Such collisions have been called single-phonon collisions.

It follows directly from (25) that electrons can interact only with long wavelength phonons whose wave vectors satisfy the inequality

$$q \leqslant k. \qquad \ldots\ldots(26)$$

The probability w of this process is proportional to the number of long wavelength phonons N_q, i.e. $w_1 \sim N_q$, and the relaxation time and mean free path are inversely proportional to w, i.e. $l \sim 1/N_q$. In both an atomic and an ionic lattice (for temperatures above the Debye temperature) the number of such phonons is proportional to the absolute temperature, and consequently $\Phi(T) \sim 1/T$.

Taking into account all that has been said above, equation (24) for the dependence of the electron mean free path on the temperature and the energy becomes

$$l \sim \epsilon^r / T \qquad \ldots\ldots(27)$$

and correspondingly the mobility

$$U = \frac{e}{m}\frac{l}{v} \sim T^{-1}\,\epsilon^{r-1/2}. \qquad \ldots\ldots(28)$$

For a non-degenerate electron gas $\epsilon \sim T$, and consequently

$$U \sim T^{r-3/2}. \qquad \ldots\ldots(29)$$

If the law $U \sim T^{-3/2}$ can be compared with (29), it follows that $r = -1$, i.e. the mean free path of electrons is inversely proportional to their energy, which contradicts the generally accepted theoretical representations, according to which l is either independent of energy (an atomic lattice, neutral impurity scattering), or increases with energy (an ionic lattice, scattering by ionic impurities). The question arises as to whether the theoretical deduction (27) that $\Phi(T) \sim T^{-1}$ is correct.

An investigation of the effect of degeneracy on mobility enables us to answer this question.

In fact, if we take a series of specimens of the same material, with different carrier concentrations and degrees of degeneracy, then at constant temperature the difference in mobility in these specimens will be due only to the variation in $\phi(\epsilon)$. On the other hand, the temperature dependence of the mobility of a specimen with a highly degenerate electron gas is completely determined by the growth in intensity of the thermal vibrations $\Phi(T)$, i.e. in this case the energy of the electrons is independent of temperature.

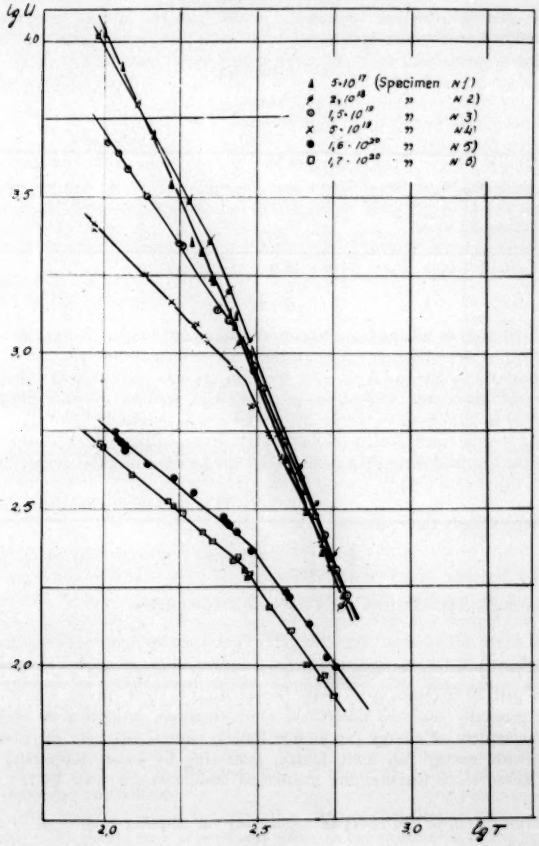


Figure 12.

Figure 12 shows the temperature dependence of the mobility for six specimens of electronic lead telluride with various carrier concentrations. For specimen No. 1 with the lowest carrier concentration $n = 5 \times 10^{17}$ and the law $U \sim T^{-3/2}$ is obeyed over the whole temperature interval. For specimen No. 2 with a carrier concentration 2×10^{18}, in which we can expect weak degeneracy at low temperatures only, $U \sim T^{-1.8}$, and at high temperatures $U \sim T^{-3/2}$. In specimen No. 3

with a carrier concentration of $1{\cdot}5 \times 10^{19}$, the slope remains at $-5/2$ for high temperatures, but becomes $-3/2$ for low temperatures. The point at which the curve deviates from a straight line is displaced to higher temperatures. In specimen No. 4 with a concentration $n = 5 \times 10^{19}$ the electron gas is degenerate for the whole temperature range over which measurements are made. At low temperatures $U \sim T^{-1}$, and $U \sim T^{-2}$ at high temperatures.

As we have already noted, the temperature dependence of the mobility for highly degenerate specimens is completely determined by the growth in intensity of the thermal vibrations, i.e. $\phi(T)$ in (24). Consequently the data shown in figure 12 show that in the high-temperature region $\Phi(T) \sim T^{-2}$, and in the low-temperature region $\Phi(T) \sim T^{-1}$.

For non-degenerate specimens $U \sim T^{-3/2}$ at high temperatures and $U \sim T^{-3/2}$ at low temperatures, i.e. the temperature dependence of the electron energy (in non-degenerate specimens) adds only another half to the degree of the temperature dependence of the mobility.

We note that $U = \tau e/m$, where τ is the relaxation time $\tau = l/v$, and consequently the additional $\frac{1}{2}$ for non-degenerate specimens is a result of the variation of v in the denominator of the expression for τ.

Thus the data obtained show that in PbTe the mean free path is independent of energy, and the anomalous temperature dependence of the mobility is due to the fact that at high temperatures $\Phi(T) \sim T^{-2}$. Analogous results were obtained for n- and p-type bismuth telluride (figure 13) and for n-type bismuth selenide. Thus the law $U \sim T^{-3/2}$ is a consequence of the fact that in the high-temperature region the mean free path of an electron is proportional to T^{-2}. This contradicts formula (27), which was deduced from perturbation theory, and shows that in this case it, or at least its first approximation (26*a*), is inapplicable. As we have already noted, it follows from (26*a*) that the scattering probability w_1 is proportional to the number of long wavelength phonons, and the latter is proportional to the temperature. But $l \sim 1/w$ and because of this we obtain the first power of the temperature in the denominator of (27). The fact that $\Phi(T) \sim T^{-2}$ at high temperatures shows that in this case the first approximation (26*a*) is not sufficient, and the amplitudes of the vibrations of the atoms are too great, and if we use perturbation theory we must also take into account the subsequent terms in the expansion of the perturbed potential. It is possible that the second term may give the transition to $w_2 \sim T^2$ and $l \sim T^{-2}$; if this is so, the matrix elements of the perturbing potential must contain terms which are the product of the amplitudes of different normal vibrations, or, what is the same thing, the product of the numbers of phonons N_{q1}, N_{q2}. The probability w_2 of such scattering events, in which not one, but two phonons of different kinds take part, must be proportional to the square of the temperature. If we take into account both single-phonon and two-phonon processes, then

$$\frac{1}{\tau} = w_1 + w_2 = aT + bT^2. \qquad \ldots\ldots(30)$$

At low temperatures where $w_1 \gg w_2$ single-phonon processes must play the decisive role and $l \sim T^{-1}$; at high temperatures $w_2 \gg w_1$ and $l \sim T^{-2}$ (see figure 14). The probability of a triple collision (one electron and two phonons) must always be much less

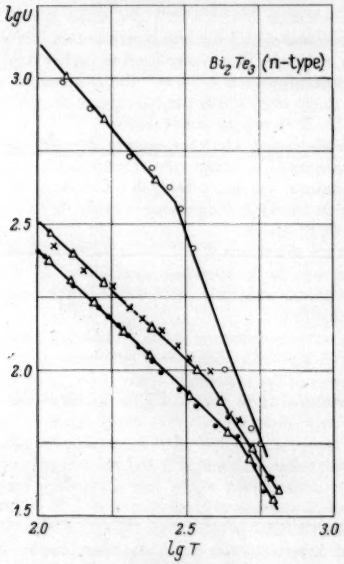


Figure 13.

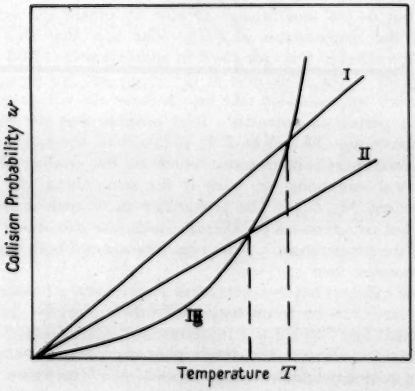


Figure 14.

than that of a double collision. However the conservation laws (25) place a very stringent condition (26) on the momenta of phonons which can take part in ' double collisions '. The number of long wavelength phonons which satisfy (26) is several orders of magnitude less than the number of short wavelength phonons. This limitation is removed for triple collisions in which two phonons take part, as the phonons can exchange momenta. Therefore, although the probability of an electron meeting two phonons is considerably less than that of a single-phonon collision, the probability of two-phonon scattering may exceed w_1 for the relatively large concentration of short wavelength phonons. By taking into account three-phonon collisions one does not introduce any new possibilities and therefore it is unlikely that their importance increases rapidly. However, experimentally r continues to increase with rise in temperature, and the mobility drops more and more quickly, its dependence going to $T^{-7/2}$ and even beyond.

From the above considerations the reason for the displacement of the break in the straight lines in figure 12 to the region of higher temperatures as the carrier concentration is increased becomes understandable. The greater the degree of degeneracy, and consequently the larger the average velocity of the electrons, the less stringent is the condition (26) $q \leqslant 2k$; a greater and greater number of phonons can take part in single-phonon collisions ; the coefficient a in (30) increases and the straight lines I and II in figure 14 get steeper and steeper, intersecting with the parabola III at higher and higher temperatures.

All the above considerations are only of a qualitative nature. They do not explain a number of facts, in particular why the slope of the straight lines for the last two specimens Nos. 5 and 6 in figure 12 is even less than that for No. 4. There is no one undisputed theory of the interaction between electrons and thermal vibrations which can explain all the experimental facts described above.

6.2. *The Scattering of Electrons in Solid Solutions*

As has already been noted the use of solid solutions of isomorphous compounds is an effective method of increasing the ratio U/κ_{ph}. It is interesting to elucidate for this case the dependence of the mean free path or the relaxation time of the electrons on their energy. It would seem that the answer to this question could be easily obtained by comparing the temperature dependence of the mobility of the pure substance and of the solid solution. Actually for a pure substance

$$U_1 = \frac{e}{m}\tau_T \quad \text{or} \quad \frac{1}{U_1} = \frac{m}{e}v_T$$

where τ_T is the relaxation time, and v_T is the number of collisions per second for scattering by thermal vibrations.

For a substance with a certain impurity content

$$\frac{1}{U_2} \simeq \frac{m}{e}(v_T + v_I)$$

where v_I is the number of collisions with impurity atoms ; consequently

$$v_I = \frac{e}{m}\left(\frac{1}{U_2} - \frac{1}{U_1}\right).$$

According to theory (Erginsoy) for neutral impurities v_I does not depend on the electron energies and is thus independent of the temperature. Figure 15 shows the temperature dependence of the electron mobilities for n-type lead telluride and for a series of PbTe–PbSe solid solutions. From these results τ_I was calculated by the method shown above ; and it was shown that v_I is approximately proportional to the temperature. If we assume, as is usually done, that the thermal vibrations of the atoms have no effect on v_I and that the temperature dependence of the ' impurity mobility ' is completely determined by the temperature dependence of the electron energy, then it follows that $l \sim \epsilon^{-1/2}$ which contradicts the generally accepted conceptions.

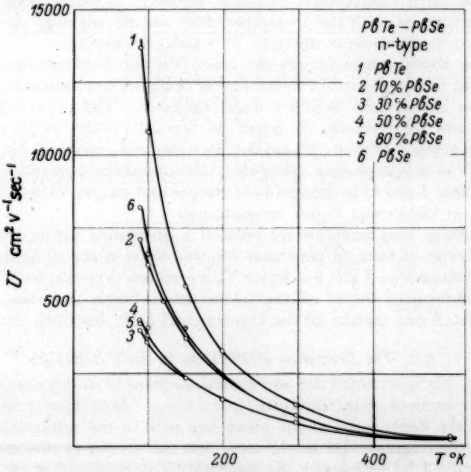


Figure 15.

To check these results analogous measurements were made on degenerate specimens, and here too it turned out that τ_I was approximately proportional to the temperature, i.e. the increase in v_I with temperature is almost independent of the increase in energy of the electrons, but is mainly due to the increase in the intensity of thermal vibrations.

Further investigations should bring out the reason for this effect, which is incomprehensible from the point of view of present-day theoretical conceptions.

6.3. *The Investigation of the Mechanism of Scattering by Ionized Impurities*

As has been noted above the thermal e.m.f. for a non-degenerate semiconductor is given by the expression

$$\alpha = \pm \frac{k}{e}[r+2-\mu^*] \qquad \ldots\ldots(31)$$

where r is the exponent in the dependence of the electron mean free path on the energy. As was shown above, the mean free path is independent of energy for scattering by thermal vibrations of the lattice, i.e. $r = 0$; for scattering by ionized impurities $l \sim \epsilon^2$, i.e. $r = 2$.

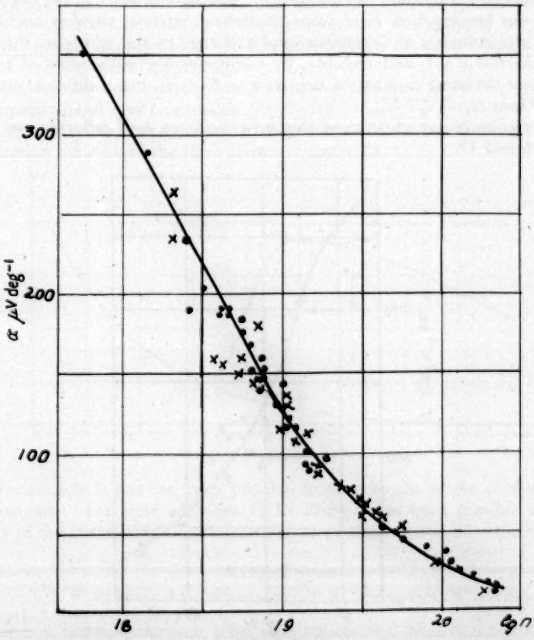


Figure 16.

It would therefore be expected that the introduction of a sufficient amount of impurity would increase the first term in (31) by two, which for the same carrier concentration corresponds to an increase in the thermal e.m.f. by

$$\frac{k}{e} \times 2 = 86 \times 2 = 172 \mu \text{v deg}^{-1}.$$

It is understood that this will reduce the carrier mobility and correspondingly the electrical conductivity of the material. However calculation shows that the introduction of a known quantity of impurity should lead to a considerable increase in the value of $\alpha^2\sigma$. To maintain the optimum carrier concentration for which $\alpha^2\sigma$ reaches its maximum value (see (10)) during this process it is necessary to

introduce supplementary donors and acceptors in equal amounts, in such a way that the two types of impurity introduced should compensate each other, leaving the carrier concentration n unchanged and equal to $N_d - N_a$ (where N_d and N_a are the numbers of donors and acceptors after the additional introduction of impurity), and increasing the number of ions to $N_i = N_d + N_a$.

Such an investigation is of some theoretical interest, since it enables us to separate the influence of degeneracy and a change in the scattering mechanisms on the thermal e.m.f. and mobility, by comparing the parameters of specimens which have the same number of carriers $n = N_d - N_a$, but a different number of impurity ions $N_i = N_d + N_a$.

The results of experiments of this type made on lead telluride are given in figures 16 and 17.

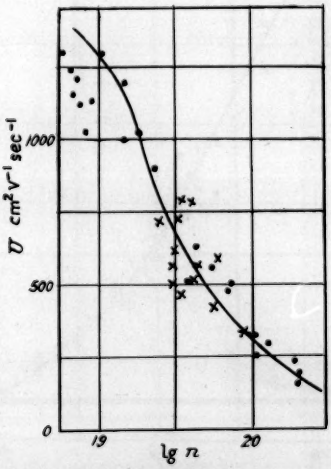


Figure 17.

Figure 16 shows the dependence of the thermal e.m.f. on the carrier concentration—the crosses show values of the thermal e.m.f. for compensated impurities, the dots show the results for uncompensated impurities. Figure 17 shows the corresponding data for the mobility. The large scatter of the points is explained by the fact that the diagram shows results from specimens obtained from different batches of tellurium ; in several cases the tellurium has been further purified by various methods (sublimation or zone melting) whilst in others it has been merely remelted. The fact that various specimens have been alloyed with different impurities also has some effect.

Judging from the fact that the points for both compensated and uncompensated specimens lie on the same curves, we may conclude that here there is no change

in the scattering mechanism. This divergence between the experimental data and theory, according to which, at such high concentrations of ionic impurities, their influence on the mobility should be very great, can evidently be explained in the following way.

1. When the distance between impurity ions is less than the electron wavelength we must consider the interaction of the electrons with the impurity lattice, rather than with separate impurity ions. In this case only the deviations from an ordered lattice of ions will cause scattering.

2. For impurity concentrations of the order of 10^{19} cm^{-3} in PbTe the formation of an impurity band may be expected.

Taking the above into account, the dependence of the mobility on the carrier concentration must be of the form shown in figure 18.

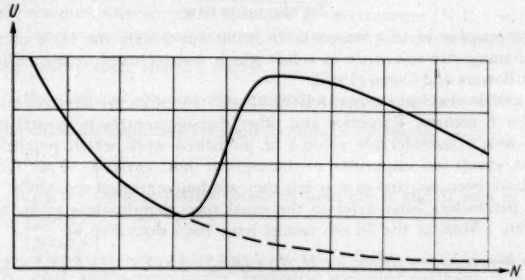


Figure 18.

Unfortunately it has not been possible to observe the whole of this curve in one substance. For solid solutions PbTe–PbSe it has been possible to observe the part of the curve which corresponds to large electron concentrations.

References

Ayrapetyans, S. V., Efimova, B. A., Stavitskaya, T. S., Stil'bans, L. S., and Sysoeva, L. U., 1957, *J. Tech. Phys.*, **27**, 2167.

Davydov, B. J., and Shmushkevich, J. M., 1940, *Uspechy Phys. Nauk* (U.S.S.R.) (Successes of Physical Sciences), **24**, 1, 21.

Devyatkova, E. D., Moiges, B. J., and Smirnov, I. A., 1959, *J. Tech. Phys.*, in the press.

Frederikse, H. P. R., 1953, *Phys. Rev.*, **91**, 431, **92**, 248.

Geballe, J. H., 1953, *Phys. Rev.*, **92**, 857.

Geballe, J. H., and Hull, G. W., 1954, *Phys. Rev.*, **94**, 1134.

Gerstein, E. Z., Stavitskaya, T. S., and Stil'bans, L. S., 1957, *J. Tech. Phys.*, **27**, 2472.

Gurevich, L. E., 1945, *J. Phys. Moscow*, **9**, 477.

Herring, C., 1953, *Phys. Rev.*, **92**, 857; 1954, *Ibid.*, **95**, 954, **96**, 1163.

Joffe, A. F., 1956, *Canad. J. Phys.*, **34**, 1342; 1957, *Semiconductor, Thermoelements and Thermoelectric Cooling* (London: Infosearch).

Joffe, A. F., Joffe, A. V., Ayrapetyans, S. V., Colomiez, N. V., and Stil'bans, L. S., 1956, *Reports Acad. of Sciences*, **106**, 6, 981.

Picus, G. E., 1951, *J. Exp. Theor. Phys.*, **21**, 852.

Price, P. I., 1955, *Phil. Mag.*, **46**, 1252; 1956, *Phys. Rev.*, **102**, 1245.

Sinany, S. S., Gordyacova, G. N., and Cocosh, G. V., 1958, *J. Tech. Phys.*, **28**, 4.

PARAMAGNETIC RESONANCE DATA

By J. W. ORTON
Clarendon Laboratory, Oxford

Abstract. A table of experimental paramagnetic resonance data is presented. Its purpose is to bring up to date (to the end of 1958) the results given in an earlier Progress Report (1955) by Bowers and Owen, and it follows closely the form of this Report. Results are given, in the form of values for the parameters in the spin Hamiltonian, for transition group ions in solids, F centres, V centres and other damage centres. No results on paramagnetic resonance in liquids or free radicals are included.

Introduction

The purpose of this Report is to bring up to date the table of results on paramagnetic resonance in solids which was included in the earlier Report of Bowers and Owen (1955).

The choice of subject matter follows closely that of its predecessor, but includes results on F centres, V centres and other damage centres in crystal lattices, as there is now a considerable volume of published work which parallels that on transition group ion impurities in diamagnetic host crystals. Free radicals and liquids have been omitted as it is felt they are better treated separately.

The parameters listed refer to the usual spin Hamiltonians given by Bowers and Owen. Most of the 3d ion results have been described by

$$\begin{aligned}\mathscr{H} = {} & \boldsymbol{\beta}(g_x H_x S_x + g_y H_y S_y + g_z H_z S_z) + D[S_z^2 - \tfrac{1}{3}S(S+1)] + E[S_x^2 - S_y^2] \\ & + A_x S_x I_x + A_y S_y I_y + A_z S_z I_z + P[I_z^2 - \tfrac{1}{3}I(I+1)] + P'[I_x^2 - I_y^2] \\ & - g_N \boldsymbol{\beta}_N \mathbf{H}.\mathbf{I}\end{aligned}$$

or, where axial symmetry holds, by

$$\begin{aligned}\mathscr{H} = {} & g_1 \boldsymbol{\beta} H_z S_z + g_\perp \boldsymbol{\beta}(H_x S_x + H_y S_y) + D[S_z^2 - \tfrac{1}{3}S(S+1)] \\ & + A S_z I_z + B(S_x I_x + S_y I_y) + P[I_z^2 - \tfrac{1}{3}I(I+1)] - g_N \boldsymbol{\beta}_N \mathbf{H}.\mathbf{I}.\end{aligned}$$

In some cases, a cubic field term may also be required which takes the form

$$\tfrac{1}{6}a[S_x^4 + S_y^4 + S_z^4 - \tfrac{1}{5}S(S+1)(3S^2+3S-1)].$$

A more general form of the Hamiltonian has been used to describe most of the results on 4f and 5f ions and is usually written as

$$\mathscr{H} = g\boldsymbol{\beta}\mathbf{H}.\mathbf{S} + B_2^0 \mathrm{P}_2^0 + B_2^2 \mathrm{P}_2^2 + B_4^0 \mathrm{P}_4^0 + B_6^0 \mathrm{P}_6^0 + B_6^6 \mathrm{P}_6^6 + (\text{nuclear terms})$$

where the P_n^m are operators having the same transformation properties as the spherical harmonics Y_n^m. For details of these operators the reader is referred to Elliott and Stevens (1953).

In cases where two or more nuclei contribute to the hyperfine structure, the term in the Hamiltonian which represents this is usually written

$$\sum_N \mathbf{S}.A^N.\mathbf{I}^N$$

where A^N represents the contribution to the hyperfine splitting from the Nth nucleus having spin I^N (see, for example, Tinkham 1956). For many cases this may be written as the sum of isotropic and anisotropic parts

$$\sum_N [\mathbf{S}.A_s{}^N.\mathbf{I}^N + \mathbf{S}.A_\sigma{}^N.\mathbf{I}^N(3\cos^2\theta - 1)]$$

where A_s represents the s orbital interaction and A_σ includes the P_σ orbital and dipole–dipole interactions.

For general references to the subject of paramagnetic resonance see Bleaney and Stevens (1953), Bowers and Owen (1955) and Bagguley and Owen (1957).

The parameters given are as follows :

T, temperature in °K

λ, approximate wavelength of operation of spectrometer (N.B. : in the case of low frequency measurements the frequency is given instead)

g, the spectroscopic splitting factor

D, E, the axial and rhombic fine structure constants in units of $10^{-4}\,\mathrm{cm}^{-1}$

a, the cubic field splitting constant in units of $10^{-4}\,\mathrm{cm}^{-1}$

$b_2{}^0 = 3B_2{}^0$
$b_2{}^2 = 3B_2{}^2$
$b_4{}^0 = 60B_4{}^0$
$b_6{}^0 = 1260B_6{}^0$
$b_6{}^6 = 1260B_6{}^6$
} fine structure constants used for 4f and 5f ions in units of $10^{-4}\,\mathrm{cm}^{-1}$

A, B, axial and perpendicular hyperfine structure constants in units of $10^{-4}\,\mathrm{cm}^{-1}$

P, P', axial and rhombic quadrupole constants in units of $10^{-4}\,\mathrm{cm}^{-1}$

ΔH, the full line width in gauss

Where signs have been determined they are given. Absence of sign indicates that it has not been determined. At the top of each sub-table the effective electronic spin S' and nuclear spins I of relevant isotopes are given for each ion. In one or two cases where a complex has a value of S' different from that specified, a note to the effect will be found in the 'remarks' column. The letter p in the 'formula' column indicates that the measurements were made on a powdered sample.

The quantities μ_N and Q which are referred to are the nuclear magnetic moment and quadrupole moment, respectively.

Acknowledgments

I should like to thank Dr. J. Owen for his constant help and encouragement in the preparation of this report. My thanks are also due to Professor B. Bleaney and Drs. D. M. S. Bagguley and J. H. E. Griffiths for many helpful suggestions, and to all those authors who have supplied details of unpublished work. The work has been done while in receipt of a maintenance grant from the Department of Scientific and Industrial Research.

$3d^3\ V^{2+}$ $S' = 3/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
MgO	$\sim 5 : 10^6$	290		$1{\cdot}9803 \pm 0{\cdot}0005$	
Vanadous phthalocyanine (p)		290	3	2·0	
		20	1·2		
$ZnSiF_6.6H_2O$		20	3	$g_\parallel = 1{\cdot}970$ $g_\perp = 1{\cdot}976$	$D = 804$

$3d^3\ Cr^{3+}$ $S' = 3/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$(NH_4)Al(SO_4)_2.12H_2O$	1 : 17	290	9	$1{\cdot}976 \pm 0{\cdot}002$	$D = 490 \pm 5$ (1 : 47)
	1 : 47	77	3		$D = 492 \pm 5$ (1 : 17)
			1·2		(at 25°c)
$(NH_4)Cr(SO_4)_2.12H_2O$	Undiluted	290	9		
		77	3	$1{\cdot}988 \pm 0{\cdot}002$	$D = 670 \pm 10$
			1·2		
$(CH_3NH_3)Al(SO_4)_2.12H_2O$	1 : 100	90	1·2	$1{\cdot}977 \pm 0{\cdot}003$	$D = -958 \pm 4$
		20			$E = -92 \pm 8$
$(CH_3NH_3)Cr(SO_4)_2.12H_2O$	Undiluted	90	1·2	$1{\cdot}976 \pm 0{\cdot}007$	$D = 871 \pm 7$
		20			$E = 92 \pm 8$
MgO	$1 : 10^3$	290	3	$1{\cdot}9800 \pm 0{\cdot}0005$	
	$1 : 10^4$	70	1·2		
	$\sim 1 : 10^5$	290	3	$g_\parallel = g_\perp = 1{\cdot}9782$	$D = 819{\cdot}4$
		70	1·2		
Al_2O_3	1 : 100			$g_\parallel = 1{\cdot}9894 \pm 0{\cdot}0010$	
	to	290	3		$D = 1920$
	$1 : 10^4$			$g_\perp = 1{\cdot}9867 \pm 0{\cdot}0010$	
CaF_2	$\sim 1 : 10^6$	90	3		
		20	1·2		
ZnF_2	—	290	—	(*a*) $g_\parallel = 1{\cdot}976 \pm 0{\cdot}007$	$D_{max} = 6020 \pm 80$
				$g_\perp = 1{\cdot}958 \pm 0{\cdot}004$	
				(*b*) $g_\parallel = 1{\cdot}975 \pm 0{\cdot}007$	$D_{max} = 5810 \pm 80$
				$g_\perp = 1{\cdot}959 \pm 0{\cdot}006$	
$CrCl_3$ (p)	Undiluted	290	3	2·00 (290°K)	
	to			2·00 (90°K)	
	13·15			2·09 (20·4°K)	
				2·37 (13·15°K)	
Aluminium acetylacetonate	1 : 50	290	3		$D = 5920 \pm 20$
			1·2	$1{\cdot}983 \pm 0{\cdot}002$	
Chromium acetylacetonate	Undiluted	290	3		$E = 520 \pm 20$
			1·2		

$I^{51}(99{\cdot}8\%) = 7/2$

H.F.S.	Remarks	Ref.
$A = 74{\cdot}3 \pm 0{\cdot}2$		Lo1 56a
		In 54
$A = -83{\cdot}9$		Ba2 u

$I^{53}(9{\cdot}5\%) = 3/2$

H.F.S.	Remarks	Ref.
	$\partial D/\partial T \simeq 2{\cdot}5 \times 10^{-4}$ cm^{-1} deg^{-1}. Approximately constant from 77°K to 290°K	Da 57
	The zero field splitting δ was also measured directly : $\delta = 1939 \pm 6$ $2(D^2+E^2)^{1/2} = 1942 \pm 10$ $\delta = 1790 \pm 30$ $2(D^2+E^2)^{1/2} = 1770 \pm 20$	Ba1 56
$A = 16{\cdot}2 \pm 0{\cdot}3$	Single line spectrum : indicates accurately cubic field	Wer1 57 Lo1 57b
$A = B = 16{\cdot}3 \pm 0{\cdot}3$	Axial field : 100 axis. Ref. Wer1 57 also reports third spectrum with 110 axis	
$A = B = 17{\cdot}0 \pm 0{\cdot}5$		Ma1 56b Geu 56 Za 56 Ma1 55
$A \sim 10$ ^{19}F h.f.s. $A_F \sim 20$ gauss	Zero field absorption occurs at $\bar{\nu} = (2847 \pm 2) \times 10^{-4}$ cm^{-1}	Ba1 u
No ^{19}F h.f.s.	$D_{max} = D_1 = D_x$. Charge compensating ion in n.n. cation site	Tin1 56
No ^{19}F h.f.s.	$D_{max} = D_1$. Axis in yz or xz plane $\pm 9°$ from xy plane. Charge compensating ion in n.n.n. cation site	
	The low temperature result may not be a true g value	Ti 55 Le 56
		Si 55

$3d^5$ Mn^{2+} $S' = 5/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
SrS–SrSe (p)		290 90 20	3	2·00	
$SrCl_2$ (p)	1 : 10^3	20	3	2	
CdTe (p)	1 : 10^4	290	3 1·2	2·008 ± 0·0025	
ZnSe (p)	5 : 10^4			2·006 ± 0·0025	
CdS (p)	1 : 10^3			2·004 ± 0·0025	
ZnS (p)	5 : 10^5			2·004 ± 0·0025	
$ZnO.Al_2O_3$ (p)	1 : 10^4			2·004 ± 0·0025	
$MgAl_2O_4$ (p)	1 : 10^3			2·004 ± 0·0025	
$6MgO.As_2O_5$ (p)	1 : 10^3			2·004 ± 0·0025	
$2MgO.\frac{2}{3}Al_2O_3$ (p)	1 : 10^3			2·004 ± 0·0025	
CaO (p)	—			2·004 ± 0·0025	
$CsCaF_3$ (p)	1 : 10^3			2·004 ± 0·0025	
CaF_2 (p)	1 : 10^3			2·004 ± 0·0025	
$KMgF_3$ (p)	1 : 10^3			2·004 ± 0·0025	
ZnS (hexagonal)	~2 : 100	290	3	2·0016 ± 0·0001	$D = -105 \pm 2$ $F = -7{\cdot}6 \pm 1{\cdot}0$
ZnS (cubic)	1 : 10^3 to 1 : 10^4		3	2·0025 ± 0·0002	$a = +7{\cdot}8$
MgO	1 : 10^3 to 1 : 10^5	290 77 4	3 1·2	2·0014 ± 0·0005	$a = +18{\cdot}65 \pm 0{\cdot}3$
CaF_2	5 : 10^3 to 4 : 10^4	290 4	3 1·2	2·001 ± 0·001	
CaF_2	~1 : 10^5	90	3	1·998 ± 0·003	$a = +0{\cdot}6 \pm 0{\cdot}4$
ZnF_2	5 : 10^6			2·002 ± 0·005	$D = -186 \pm 8$ $E = -41 \pm 8$
$CdCl_2$		290 90 20	1·2	2·00	$D = 15 \pm 5$
LiCl				2·003 ± 0·001	
LiF					

$I^{55}(100\%) = 5/2$

H.F.S.	Remarks	Ref.
$A_1 = 89$ $A_2 = 86$	Two sets of six lines. May be due to S and Se neighbours. Cf. ZnS and ZnSe	Lo1 55
$A^{55} = 87{\cdot}4$ $A^{53} = 91{\cdot}0$	$I(^{53}\text{Mn}) = 7/2 = \mu_N{}^{53}/\mu_N{}^{55} = 1{\cdot}455 \pm 0{\cdot}002$	Do2 56b
$A = 55 \pm 1$ $A = 61 \pm 1$ $A = 65 \pm 1$ $A = 65 \pm 1$ $A = 76 \pm 1$ $A = 81 \pm 1$ $A = 81 \pm 1$ $A = 81 \pm 1$ $A = 85 \pm 1$ $A = 91 \pm 1$ $A = 93 \pm 1$ $A = 92 \pm 1$		Va 55
$A = -65 \pm 1$	$\mathscr{H} = g\beta\mathbf{H}.\mathbf{S} + D[S_z{}^2 - \frac{1}{3}S(S+1)]$ $+ (F/180)\,[35S_z{}^4 - 30S(S+1)\,S_z{}^2 + 25S_z{}^2 - 6S(S+1)$ $+ 3S^2(S+1)^2] + A\mathbf{I}.\mathbf{S}$	Kel 58
$A = 64{\cdot}0 \pm 0{\cdot}1$	Cubic field	Mat 55 Mat 56 Wat 58
$A = -81{\cdot}0 \pm 0{\cdot}2$	Cubic field	Lo1 56a Lo1 57a
$A = 95 \pm 1$	Line width 40 gauss probably due to ^{19}F h.f.s.	Lo1 57a
$A = 97{\cdot}8 \pm 1{\cdot}0$ $A_a = 9{\cdot}5 \pm 0{\cdot}3$ $A_e = 2{\cdot}7 \pm 0{\cdot}5$	A_a and A_e represent the interaction with neighbouring ^{19}F nuclei	Ba4 58
$A = -96 \pm 3$ $A_y{}^{\text{I}} = 16{\cdot}5 \pm 0{\cdot}7$ $A_z{}^{\text{I}} = 18{\cdot}2 \pm 0{\cdot}2$ $A_y{}^{\text{II}} = 14{\cdot}6 \pm 1{\cdot}2$ $A_z{}^{\text{II}} = 12{\cdot}5 \pm 0{\cdot}2$	I and II represent inequivalent F^- ions	Tin1 56
$A = 82 \pm 5$	Trigonal distortion on cubic field	Pa u
$A = 76 \pm 5$	Crystals grown from melt	Fu 58
No h.f.s.	(*a*) Annealed—two lines (*b*) Quenched—one line } Crystals grown from melt	Fu 58

$3d^5$ Mn^{2+}

Formula	Dilution	T	λ (cm)	g	Fine structure
NaCl	1 : 100 to 1 : 10^6	290 80	3 1·2	(*a*) 2·015 ± 0·005 (*b*) 2·0011 ± 0·0005 (*c*) 2·004 ± 0·002 (*d*) 2·015 ± 0·005 (*e*) 2·0022 ± 0·0008	Complex spectrum
KCl		290 80	3 1·2	(*a*) 2·002 ± 0·001 (*b*) 2·020 (*c*) 2·002 ± 0·001 (*d*) 2·0045 ± 0·0005	
NaBr		290	3	(*a*) 2·07 ± 0·1 (*b*) ~2·0	
KBr KI		290 80	3 1·2	(*a*) ~2·0 (*b*) 2·0045	
NaF	~1 : 10^4			(*a*) 2·00 ± 0·01 (*b*) 1·996 ± 0·006	D_z = 150 ± 5 D_x = 59 ± 3
AgCl	1 : 10^4 to 1 : 100	290	3	(*a*) 2·006 ± 0·003 (*b*) 2·000 ± 0·003	— —
AgBr	1 : 10^4 to 1 : 100	290	3	2·006 ± 0·010	—
Manganous phthalocyanine (p)	Undiluted	290 20	3 1·2	2·0	—
$K_4Fe(CN)_6.3H_2O$	5 : 100 to 1 : 10^3	12	3 1·2	g_a = 2·624 ± 0·008 g_y = 2·182 ± 0·008 g_z = 0·63 ± 0·10	
Copper	Atomic % Mn 11·1, 5·6, 1·4, 0·5, 0·07	290 77 4·2	3 1·2	2·01 ± 0·01	
Silver	4·2 at. %	77	3	1·99 ± 0·01	
Magnesium	0·67 at. %	4	3	2·02 ± 0·01	
Germanium				2·0061 ± 0·0002	a = +8·8 ± 0·5

(*continued*)

H.F.S.	Remarks	Ref.
No h.f.s. $A = 82{\cdot}9 \pm 0{\cdot}5$ No h.f.s.	(*a*) Unannealed $\Delta H \sim 300$ gauss (*b*) Annealed $\Delta H \simeq 15$ gauss (*c*) Slow cool $\Delta H = 39 \pm 4$ gauss } Crystals grown from mixed melt	Os 55 Ab2 56 Fo 56
No h.f.s. $A = 81{\cdot}0 \pm 0{\cdot}2$	(*d*) $\Delta H = 220 \pm 20$ gauss (*e*) $\Delta H \sim 7$ gauss } Crystals grown from mixed aqueous solution	Lo1 56c
No h.f.s. No h.f.s. No h.f.s. $A = 88{\cdot}9 \pm 0{\cdot}5$	(*a*) Unannealed $\Delta H \sim 60$ gauss (*b*) Quenched $\Delta H = 270 \pm 30$ gauss (*c*) Annealed $\Delta H = 47 \pm 3$ gauss } From mixed melt (*d*) $\Delta H \sim 45$ gauss. Annealing gives the single line 47 gauss wide, as above } From aqueous solution and melt	Ab2 56 Fo 56 Lo1 56c Fu 58
No h.f.s.	(*a*) Slow cool $\Delta H \sim 1400$ gauss (*b*) Quench $\Delta H \sim 3000$ gauss } From mixed melt	Ab2 56
No h.f.s. $A = 88{\cdot}9$	(*a*) Slow cool from melt $\Delta H \sim 1200$ gauss (*b*) From aqueous solution $\Delta H \simeq 45$ gauss. Annealing gives single line as for KCl	Ab2 56 Fo 56 Lo1 56c
$A = 92 \pm 4$ $A = 91 \pm 4$ $A_a = 14{\cdot}4 \pm 0{\cdot}3$ $A_o = 2{\cdot}8 \pm 0{\cdot}7$	(*a*) and (*b*) are two distinct spectra. Both show h.f.s. due to interaction with ^{19}F nuclei but the intensity of (*a*) was too small to permit analysis. A_a and A_o refer to the ^{19}F interaction. Crystals grown from melt	Ha 58
$A = 81 \pm 4$ —	$\Delta H = 19 \pm 3$ gauss $\Delta H = 52 \pm 4$ gauss } Crystals from melt	Ab1 57
$A = 78 \pm 4$	$\Delta H = 60 \pm 5$ gauss Crystals from melt	Ab1 57
—		In 54
$A_x = 84{\cdot}5 \pm 0{\cdot}5$ $A_y = 46{\cdot}5 \pm 0{\cdot}5$ $A_z = 104 \pm 20$	$S' = 1/2$ Direction cosines (*a*) x: 0·864, y: 0·410, z: 0·295 (*b*) x: ∓0·105, y: ∓0·423, z: ±0·899 (*c*) x: 0·495, y: −0·807, z: −0·322	Ba3 56
	No structure of any kind even in a single crystal. Resonance shifts to low field at low temperatures. Minimum ΔH is ~100 gauss	Ow 57
		Ow 57
		Ow 57
$A = -42{\cdot}4 \pm 0{\cdot}3$	Mn actually present as Mn^{3-}	Wat 57

$3d^5$ Fe^{3+}

Formula	Dilution	T	λ (cm)	g	Fine structure
MgO		290			
	$<1:10^3$	77	1·2	2·0037 ± 0·0007	$a = +205$
		4			
Haemoglobin, myoglobin and derivatives	Undiluted	20	3	$g_\parallel = 6{\cdot}0$	
			1·2		
	Fe ~1 in 10^4 atoms		1·0	$g_\perp = 2{\cdot}0$	
			0·8		
			0·6		

					I	II
$K_3Fe(CN)_6$	Undiluted	20	4	(*a*)	2·36	2·31
			1·2	(*b*)	2·21	2·39
				(*c*)	—	0·94

$3d^5$ Cr^+ $S' = 5/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
NaF	~1 : 10^5			2·000 ± 0·002	$a = 3{\cdot}7 \pm 0{\cdot}7$

$3d^8$

Formula	Dilution	T	λ (cm)	g	Fine structure
ZnF_2		90	4	$g_\parallel = 8{\cdot}97 \pm 0{\cdot}02$	$\Delta(20°K) = 2030 \pm 40$
	$<3:10^4$	20	1·2	$g_\perp = 0$	$\Delta(90°K) = 2240 \pm 100$
		12			
MgO		4	1·2	3·428	—
				6·90	
ZnS				2·26	
$BaTiO_3$†		290	3	(*a*) 3·52 }	
				(*b*) 2·72 } 290°K	
				(*c*) 2·41 }	
				(*d*) 1·79 }	
		>120°C		(*e*) 2·0	
		<0°C		(*f*) 2	
Iron phthalocyanine (p)	Undiluted	290	3	(*a*) 3·8	
		20	1·2	(*b*) 2·0	
NaCl‡		290	3	2·0	
KCl‡		290	3	2·0	
$CdCl_2$		20	1·2	$g_\perp = 7{\cdot}4$	

† See addendum.

$S' = 5/2$

H.F.S.	Remarks	Ref.
	Accurately cubic field. Sign of a from intensities at 4°K	Lo1 56d
	One line only. The axis of g is perpendicular to the haem plane. The g values are interpreted by assuming the ^6S ground state is split into 3 doublets with zero field splitting $\geqslant 10$ cm^{-1}	Be1 57
	I, H_{tf} along c axis ($S' = 1/2$) II, H_{tf} perpendicular to c axis	Ba3 56

$I^{53}(9{\cdot}5\%) = 3/2$

H.F.S.	Remarks	Ref.
$A = 14{\cdot}0 \pm 0{\cdot}5$ $A_\sigma = 12{\cdot}7 \pm 0{\cdot}2$ $A_\sigma = 0{\cdot}9 \pm 0{\cdot}8$	Produced by irradiation of Cr^{3+} in NaF with x-rays, γ-rays, or electrons. A_σ and A_σ refer to interaction with ^{19}F nuclei	Ha 58

Fe^{2+}

H.F.S.	Remarks	Ref.
$A_x^{I} = 23{\cdot}9 \pm 1{\cdot}2$ gauss $A_x^{II} = 15{\cdot}3 \pm 1{\cdot}0$ gauss Mean $A =$ 22×10^{-4} cm^{-1}	I and II represent inequivalent F^- ions. The spin Hamiltonian used is $\mathcal{H} = g_x\beta S_x H_x + \Delta S_z$ ($S' = 1/2$). The 'mean A' in cm^{-1} is calculated from the true spin formalism (see Pt. II of the ref.)	Tin1 56
—	The $g = 6{\cdot}9$ line is a double quantum jump. $S' = 1$ in octahedral field	Lo1 56a
	$S' = 2$ in tetrahedral field	Lo1 56a
	The iron was added as Fe_2O_3 but resonances ascribed to Fe^{2+} on account of large crystal field splittings. There is no evidence given that these are true g values	Hor 54
	(e) One line only (f) Two lines	In 54
	$\Delta H = 290 \pm 40$ gauss	Ab2 56
	$\Delta H = 400 \pm 40$ gauss	Ab2 56
	Line shape asymmetrical, characteristic of non-Kramers doublet ($S' = 1/2$)	Pa u

‡ Valency state not specified.

$3d^7\ Co^{2+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
MgO	$\sim 1:10^4$	20	3	$4{\cdot}278 \pm 0{\cdot}001$	
ZnF_2		20		$g_z = (2{\cdot}6)$ $g_y = 6{\cdot}05 \pm 0{\cdot}01$ $g_x = 4{\cdot}1 \pm 0{\cdot}1$	
CaF_2	$\sim 1:10^5$	20	3	$g_\parallel \sim 6{\cdot}6$ $g_\perp = 3{\cdot}40 \pm 0{\cdot}05$	
NaF	$\sim 1:10^5$	20		$g_x = 4{\cdot}3$ $g_y = 3{\cdot}3$ $g_z = 5{\cdot}7$	
$CdCl_2$		20	1·2	$g_\parallel = 3{\cdot}06 \pm 0{\cdot}02$ $g_\perp = 4{\cdot}98 \pm 0{\cdot}02$	
$ZnK_2(SO_4)_2.6D_2O$	$^{58}Co \sim 1:10^7$ $^{57}Co \sim 1:10^7$	20	3		
$ZnK_2(SO_4)_2.6D_2O$	~ 4 mc ^{58}Co	20	3		
$Zn(NH_4)_2(SO_4)_2.6D_2O$	$^{59}Co \sim 1:10^3$ $^{60}Co \sim 1:10^5$	20 4	3	$g_\parallel = 6{\cdot}41$	
Cobalt phthalocyanine	Undiluted	290 20	3 1·2	(*a*) 2·9 (*b*) 1·98–2·4	

$3d^7\ Fe^+$

Formula	Dilution	T	λ (cm)	g	Fine structure
MgO	$\sim 1:10^3$	20	3	$4{\cdot}15 \pm 0{\cdot}01$	
NaF	$\sim 1:10^4$	20 4		$4{\cdot}344 \pm 0{\cdot}002$	

$3d^8\ Ni^{2+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
MgO		290 77 4		$2{\cdot}225 \pm 0{\cdot}005$ (290) $2{\cdot}227 \pm 0{\cdot}002$ (77) $2{\cdot}234 \pm 0{\cdot}002$ (4)	
$NiSiF_6.6H_2O$	Undiluted	290		$2{\cdot}34 \pm 0{\cdot}02$	$D = 0$ at 6200 kg cm^{-2} $\partial D/\partial P = +0{\cdot}834 \times 10^{-4}$ cm^{-1}(kg cm^{-2})$^{-1}$ $1/D(\partial D/\partial u) = 3{\cdot}18 \times 10^{-4}$ (kg cm^{-2})$^{-1}$

$I^{59}(100\%) = 7/2$

H.F.S.	Remarks	Ref.
$A = 97{\cdot}8 \pm 0{\cdot}2$	Accurately cubic field	Lo1 58b
$A_x = (-43)$ $A_y = 217 \pm 2$ $A_z = (67)$ $A_y^F = 32 \pm 1$ $A_z^F = 21 \pm 5$	Values in parentheses are calculated theoretically from data available A^F is the h.f.s. constant for the ^{19}F nuclei and is a mean value for the two inequivalent F^- ions	Tin1 56
$B = 170 \pm 4$ ^{19}F h.f.s. $A_F \sim 30$		Ba1 u
$A_y = 82$ $A_z = 250$ ^{19}F h.f.s.	^{19}F h.f.s. not analysed	Ha 58
$A = 35$ $B = 170$	Trigonal distortion to cubic field	Pa u
	$\mu_N^{58}/\mu_N^{59} = 0{\cdot}829 \pm 0{\cdot}002$ $I(^{58}\text{Co}) = 4$	Jo 56 Ba5 56
	$\mu_N^{58}/\mu_N^{59} = 0{\cdot}8734 \pm 0{\cdot}0024$	Do1 57
$A^{59} = 252$ $A^{60} = 144$	$\mu_N^{60}/\mu_N^{59} = 0{\cdot}8191 \pm 0{\cdot}0016$ $I(^{60}\text{Co}) = 5$	Do2 56a
		In 54

$S' = 1/2$

H.F.S.	Remarks	Ref.
	Produced by irradiation of Fe^{2+} with x-rays	Or u
Fluorine h.f.s. $A_s = 28{\cdot}3 \pm 1{\cdot}0$ $A_o = 10 \pm 1$	Produced by irradiation with ^{60}Co γ-rays, electrons, x-rays A_s and A_o represent interaction with neighbouring ^{19}F nuclei	Bl1 57

$S' = 1$

H.F.S.	Remarks	Ref.
		Lo1 56a Lo1 58a
	Variation of fine structure with pressure (P) from 1 to 1000 atmospheres. Also variation with axial stress (U). $\partial g/\partial P = 0$	Wa2 57 Wa1 58

$3d^8$ Ni^{2+}

Formula	Dilution	T	λ (cm)	g	Fine structure
$NiCl_2$ (p)		290 to 20	3	2·30 (20·4°K) 2·46 (40·6°K) 2·27 (50·2°K) 2·24 (90°K) 2·25 (290°K)	
$CdCl_2$		20	1·2	2·28	$D = 14100$
Germanium		20	2	2·109	

$3d^8$ Co^+ $S' = 1$

Formula	Dilution	T	λ (cm)	g	Fine structure
NaF		90		$2·31 \pm 0·02$	

$3d^9$ Cu^{2+} $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$Cu(NH_3)_4SO_4.H_2O$	Undiluted	290	3 1·2	$g_\parallel = 2·22 \pm 0·02$ $g_\perp = 2·05 \pm 0·02$	
$Cu(NH_3)_4(NO_3)_2$	Undiluted	290	1·2	(*a*) $2·07 \pm 0·02$ (*b*) $2·14 \pm 0·02$ (*c*) $2·02 \pm 0·02$	
$CuCl_2.2H_2O$	Undiluted	77 20	0·9	(*a*) 2·187 (*b*) 2·037 (*c*) 2·252	
$ZnK_2(SO_4)_2.6H_2O$	1 : 200	77	526 Mc/s	$g_\parallel = 2·47$ $g_\perp = 2·08$	
NaCl		290	3 cm	(*a*) 2·07 (*b*) 2·06 (*c*) 2·16	
KCl		290	3 cm	$2·180 \pm 0·009$	
Copper dimethylglyoxime	Undiluted			$g_x = 2·065$ $g_y = 2·033$ $g_z = 2·136$	
MgO (p)				$2·190 \pm 0·002$	
Copper phthalocyanine (p)	Undiluted	290 20	1·2 cm 0·8 cm	$g_\parallel = 2·165 \pm 0·004$ $g_\perp = 2·045 \pm 0·003$	

(*continued*)

H.F.S.	Remarks	Ref.
	The low temperature results may not be true g values	Le 56
	Trigonal distortion to cubic field	Pa u
$A^{61} = 9{\cdot}9$	Enriched ^{61}Ni, $I^{61} = 3/2$. Present as Ni^{3-}	Wo 58

$I^{59}(100\%) = 7/2$

H.F.S.	Remarks	Ref.
	$\Delta H \sim 200$ gauss probably due to unresolved h.f.s. Produced by irradiation of Co^{2+}	Ha 58

$I^{63}(69{\cdot}1\%) = 3/2$ $I^{65}(30{\cdot}9\%) = 3/2$

H.F.S.	Remarks	Ref.
	Exchange narrowing of lines $\Delta H \sim 40$ gauss	Ok 54 Ca 56
	Exchange narrowing	Ok 54
		Ge 55
$A = -83$ $B = 45$ $Q = 10$		Ga2 55
	(*a*) Slow cool (*b*) Quench $\Delta H \sim 2400$ gauss } From mixed melt (*c*) From aqueous solution. $\Delta H = 190 \pm 15$ gauss	Ab2 56
	Slow cool from mixed melt $\Delta H = 400 \pm 25$ gauss	Ab2 56
	Maximum g value along axis of needle	Ja 58
$A = 19 \pm 1$	Cubic field	Ha1 u
$A = 208 \pm 10$ $B = 31 \pm 10$	Axis normal to plane of N atoms	Be2 55

$3d^9\ Cu^{2+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
Copper tetra-phenyl porphin :					
(1) Unchlorinated				(1) 2·18	
(2) Chlorinated				(2) 2·18	
$CdCl_2$		290	1·2	$g_\parallel = 2{\cdot}205 \pm 0{\cdot}010$	
				$g_\perp = 2{\cdot}19 \pm 0{\cdot}01$	
		90	1·2	$g_\parallel = 2{\cdot}34 \pm 0{\cdot}01$	
		20		$g_\perp = 2{\cdot}07 \pm 0{\cdot}01$	

$3d^9\ Ni^+$

Formula	Dilution	T	λ (cm)	g	Fine structure
NaF	$\sim 1 : 10^5$	20		$g_\parallel = 2{\cdot}766 \pm 0{\cdot}002$	
				$g_\perp = 2{\cdot}114 \pm 0{\cdot}002$	

$4d^3\ Tc^{4+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
K_2PtCl_6		4	1·2	$1{\cdot}9896 \pm 0{\cdot}0005$	

$4f'\ Ce^{3+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
$La_2Mg_3(NO_3)_{12}.24H_2O$		4	3	$(g_\parallel = 0{\cdot}25)$	
				$g_\perp = 1{\cdot}84$	
CaF_2	$5 : 10^5$		3	$g_\parallel = 3{\cdot}030 \pm 0{\cdot}003$	
			1·2	$g_\perp = 1{\cdot}396 \pm 0{\cdot}002$	

$4f^2\ Pr^{3+} \quad S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$Y(C_2H_5SO_4)_3.9H_2O$				$g_s = 1{\cdot}52 \pm 0{\cdot}02$	

$4f^3\ Nd^{3+} \quad S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$La_3Mg_3(NO_3)_{12}.24H_2O$		4	3	$g_\perp = 2{\cdot}72$	

(*continued*)

H.F.S.	Remarks	Ref.
$A_{Cu} = 250$ $A_{Cl} = 120$	Electron interacts with two Cl nuclei	Be2 56
	Trigonal distortion to cubic field	Pa u
$A = 120 \pm 10$ B, P undetermined	Three ions per unit cell. Axes : crystal cubic axes Complex h.f.s. from Cl nuclei	

$S' = 1/2$

H.F.S.	Remarks	Ref.
$A_a^{I} = 41 \pm 2$ $A_a^{II} = 10 \pm 1$ $A_o^{I} = 16 \pm 2$ $A_o^{II} = 0 \pm 2$	The ^{19}F h.f.s. constants A^{I} and A^{II} refer to inequivalent F^- ions	Ha 58

$S' = 3/2 \quad I^{99} = 9/2$

H.F.S.	Remarks	Ref.
$A_{max} = 138{\cdot}3 \pm 0{\cdot}4$ $A_{min} = 133{\cdot}9 \pm 0{\cdot}4$	Another spectrum visible, possibly due to Re^{4+}	Lo2 58

$S' = 1/2$

H.F.S.	Remarks	Ref.
$B^{141} = 126 \pm 1$	g_1 quoted from previous work. $I(^{141}Ce) = 7/2$. ^{141}Ce produced artificially	Ke 57
^{19}F h.f.s.	The ^{19}F h.f.s. is attributed to nuclear ' spin flips ' due to weak coupling to the electron transitions. They are separated by the N.M.R. frequency in the same field	Ba1 u

$I^{141}(100\%) = 5/2$

H.F.S.	Remarks	Ref.
$A = 750 \pm 20$	$\mathcal{H} = g_1 \beta H_z S_z + A S_z I_z + \Delta_x S_x + \Delta_y S_y$	Ba2 55a

$I^{143}(12{\cdot}2\%) = 7/2 \quad I^{145}(8{\cdot}3\%) = 7/2$

H.F.S.	Remarks	Ref.
$B^{143} = 310$ $B^{145} = 190$ $B^{147} = 237 \pm 1$	$I(^{147}Nd) = 5/2$	Ke 57

$4f^3\ Nd^{3+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
CaF_2	$1:10^3$	20		$g_\parallel = 4{\cdot}412 \pm 0{\cdot}008$ $g_\perp = 1{\cdot}301 \pm 0{\cdot}002$	
SrF_2	1 : 100	20		$g_\parallel = 4{\cdot}289 \pm 0{\cdot}008$ $g_\perp = 1{\cdot}505 \pm 0{\cdot}002$	
$LaCl_3$			3		

$4f^7\ Eu^{2+}\quad S' = 7/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
SrS			3		
$SrCl_2$	$1:10^5$	290		$1{\cdot}994 \pm 0{\cdot}001$	$\delta = 370 \pm 70$
KCl (p)		77	3	2	
CaF_2	$1:10^4$	90	3	$1{\cdot}989 \pm 0{\cdot}002$	$b_4^0 = 57{\cdot}9 \pm 0{\cdot}2$ $b_6^0 = 0{\cdot}5 \pm 0{\cdot}2$ $\delta = 1810 \pm 50$

$4f^7\ Gd^{3+}\quad S' = 7/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$Bi_2Mg_3(NO_3)_{12}.24H_2O$	$2:10^4$	290 77	3 1·2	1·991	
$LaCl_3.7D_2O$	$2:10^4$	290 77	3 1·2	1·99	
$LaCl_3$–$CeCl_3$	$1:10^3$(Gd) 1 : 100(Ce)	290 90 20	3	$1{\cdot}991 \pm 0{\cdot}001$	90°K $b_2^0 + 16{\cdot}0 + 0{\cdot}02$ $b_4^0 + 2{\cdot}13 \pm 0{\cdot}05$ $b_6^0 + 0{\cdot}25 \pm 0{\cdot}05$ $b_6^6 + 1{\cdot}40 \pm 0{\cdot}30$ 290°K $b_2^0 + 8{\cdot}36 \pm 0{\cdot}10$ $b_4^0 + 1{\cdot}68 \pm 0{\cdot}04$ $b_6^0 + 0{\cdot}64 \pm 0{\cdot}15$
CaF_2	$\sim 1:10^4$	290	1·2	$1{\cdot}991 \pm 0{\cdot}002$	$b_4^0 = +46 \pm 1$ $b_6^0 = -1{\cdot}0 \pm 0{\cdot}5$ $\delta = 1490 \pm 50$

(*continued*)

H.F.S.	Remarks	Ref.
No ^{19}F h.f.s.	Three ions per unit cell. g axis is the cubic axis	Bl4 56
No ^{19}F h.f.s.	Three ions per unit cell. g axis is the cubic axis	Bl4 56
$B^{145} = 103{\cdot}4 \pm 0{\cdot}2$ $P^{145} = -0{\cdot}0127 \pm 0{\cdot}003$	Double resonance	Hal 58

$I^{151}(47{\cdot}8\%) = 5/2$ $I^{153}(52{\cdot}2\%) = 5/2$

H.F.S.	Remarks	Ref.
h.f.s. from ^{151}Eu, ^{152}Eu and ^{153}Eu	$I(^{152}\text{Eu}) = 3$ $\mu_N^{152} = 2{\cdot}03$ m.m. (taking $\mu_N^{151} = 3{\cdot}6$ n.m.)	Ma2 57
$A^{151} = 32{\cdot}5 \pm 0{\cdot}5$ $A^{153} = 14{\cdot}7 \pm 0{\cdot}5$	δ is the overall splitting of the states in cubic field	Lo1 56a Ba1 u
$A^{151} = 32{\cdot}56 \pm 0{\cdot}06$ $A^{152} = 15{\cdot}12 \pm 0{\cdot}15$ $A^{153} = 14{\cdot}38 \pm 0{\cdot}03$ $A^{154} = 15{\cdot}67 \pm 0{\cdot}06$	$I(^{152}\text{Eu}) = 3$, $I(^{154}\text{Eu}) = 3$ $\mu_N^{152}/\mu_N^{151} = 0{\cdot}5574 \pm 0{\cdot}006$ $\mu_N^{151}/\mu_N^{153} = 2{\cdot}264 \pm 0{\cdot}006$ $\mu_N^{152}/\mu_N^{154} = 0{\cdot}965 \pm 0{\cdot}010$	Abr3 57
$A^{151} = 34{\cdot}5 \pm 0{\cdot}2$ $A^{153} = 15{\cdot}4 \pm 0{\cdot}1$ Also ^{19}F h.f.s. due to nuclear 'spin flips'	$\mathcal{H} = g\beta\mathbf{H}.\mathbf{S} + B_4(P_4^0 + 5P_4^4) + B_6(P_6^0 - 20P_6^4) + A\mathbf{I}.\mathbf{S}$ δ is the overall cubic field splitting of the states ($= 32b_4^0 - 8b_6^0$)	Ma1 56a Ry 56 Ba4 58

$I^{155}(14{\cdot}7\%) = 3/2$ $I^{157}(15{\cdot}7\%) = 3/2$

H.F.S.	Remarks	Ref.
$A^{155} = 3{\cdot}7 \pm 0{\cdot}3$ $A^{157} = 4{\cdot}96 \pm 0{\cdot}17$	$\mu_N^{155}/\mu_N^{157} = 0{\cdot}75 \pm 0{\cdot}07$	Lo1 56b
$A^{157} = 5{\cdot}3 \pm 0{\cdot}3$	^{158}Gd h.f.s. not resolved	Lo1 56b
	Signs of parameters from intensities at 20°K. The $CeCl_3$ was added to prevent saturation	Hut2 57
	Cubic field. δ is the overall splitting of the states in cubic field. Signs of parameters from intensities at 4°K	Lo1 58c

$4f^9\ Dy^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$Y(CH_3COO)_3.4H_2O$	1 : 150	4	3	$g_z = 13{\cdot}60 \pm 0{\cdot}06$	
$La_2Mg_3(NO_3)_{12}.24H_2O$	5 : 10^4		3	$g_\parallel = 4{\cdot}281 \pm 0{\cdot}006$ $g_\perp = 8{\cdot}923 \pm 0{\cdot}16$	

$4f^{10}\ Ho^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$Y(C_2H_5SO_4)_3.9H_2O$	1 : 100	13		$g_z = 15{\cdot}410 \pm 0{\cdot}010$†	

$4f^{11}\ Er^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
CaF_2	1 : 10^5	20	3 1·2	(*a*) $6{\cdot}78 \pm 0{\cdot}01$ (*b*) $g_\parallel = 7{\cdot}76 \pm 0{\cdot}02$ $g_\perp = 6{\cdot}253 \pm 0{\cdot}006$ (*c*) $g_\parallel = 6{\cdot}76 \pm 0{\cdot}02$ $g_\perp = 9{\cdot}11 \pm 0{\cdot}01$	

$4f^{13}\ Yb^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$Y(CH_3COO)_3.4H_2O$	1 : 100	4	3	$g_z = 4{\cdot}57 \pm 0{\cdot}02$	

$5f^1\ (NpO_2)^{2+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$UO_2Rb(NO_3)_3$	2 : 10^5	4	3	$g_\parallel = 3{\cdot}40$	

NpO_2^{2+} in $UO_2Rb(NO_3)_3$. Results in

† These figures are more accurate

$I^{161}(21{\cdot}1\%) = 5/2$ $I^{163}(24{\cdot}8\%) = 5/2$

H.F.S.	Remarks	Ref.
$A_z^{161} = 381 \pm 5$ $A_z^{163} = 540 \pm 5$	$\mu_N^{163}/\mu_N^{161} = 1{\cdot}41 \pm 0{\cdot}02$	Coo 56
$A^{161} = 116{\cdot}1 \pm 0{\cdot}7$ $B^{161} = 246{\cdot}3 \pm 1{\cdot}5$ $P^{161} = 14{\cdot}2 \pm 1{\cdot}0$ $A^{163} = 162{\cdot}2 \pm 0{\cdot}7$ $B^{163} = 341{\cdot}5 \pm 1{\cdot}5$ $P^{163} = -16{\cdot}8 \pm 1{\cdot}0$	$\mu_N^{163}/\mu_N^{161} = -1{\cdot}392 \pm 0{\cdot}020$ $Q^{163}/Q^{161} = +1{\cdot}18 \pm 0{\cdot}15$	Pa 58

$I^{165}(100\%) = 7/2$

H.F.S.	Remarks	Ref.
$A = 3340 \pm 10$ $B = 200 \pm 40$ $P = 3 \pm 3$	$\mathscr{H} = g_z\beta H_z S_z + AS_z I_z + B(S_z I_z + S_y I_y)$ $+ P[I_z^2 - \frac{1}{3}I(I+1)] + \Delta_x S_x + \Delta_y S_y$	Ba2 55b

$I^{167}(22{\cdot}9\%) = 7/2$

H.F.S.	Remarks	Ref.
(*b*) $A = 261 \pm 3$ $B = 219 \pm 3$	(*a*) Single isotropic line (*b*), (*c*) Three lines from three inequivalent ions per unit cell. Axis of *g*, crystal cubic axis	Ba1 u

$I^{171}(14{\cdot}3\%) = 1/2$ $I^{173}(17{\cdot}0\%) = 5/2$

H.F.S.	Remarks	Ref.
$A_z^{171} = 1220 \pm 10$ $A_z^{173} = 341 \pm 3$	$\mu_N^{173}/\mu_N^{171} = 1{\cdot}39 \pm 0{\cdot}01$	Coo 56

$I^{237} = 5/2$ $I^{239} = 1/2$

H.F.S.	Remarks	Ref.
$A^{239} = 503 \pm 5$	$\mu_N^{237}/\mu_N^{239} = 16{\cdot}45 \pm 0{\cdot}17$	Abr2 57

ref. Abr 2 57 withdrawn (Ref. 1).
than those in the reference.

$5f^3\,U^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	1 : 100 to 1 : 10^4	20	3	$g_\parallel = 4{\cdot}153 \pm 0{\cdot}005$	
		4	1·2	$g_\perp = 1{\cdot}520 \pm 0{\cdot}002$	
CaF_2	1 : 10^3	20		$g_\parallel = 3{\cdot}501 \pm 0{\cdot}008$ $g_\perp = 1{\cdot}866 \pm 0{\cdot}002$	
SrF_2	1 : 100	20		$g_\parallel = 3{\cdot}433 \pm 0{\cdot}008$ $g_\perp = 1{\cdot}971 \pm 0{\cdot}002$	

$5f^7\,Cm^{3+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	1 : 10^3	290 77	3	1·9914 ± 0·0008	$b_2^0 = 7{\cdot}6 \pm 0{\cdot}2$ $b_4^0 = 2{\cdot}0 \pm 0{\cdot}2$ $D = 10\ (= b_2^0)$

F Centres $S' = 1/2$ $I(^{19}F) = 1/2$

Formula	No. per cm³	T	λ (cm)	g	H.F.S.		
LiF		290	3	1·999 ± 0·001	Not resolved		
LiF	~10^{17}		3		19 line h.f.s.		
LiF		4	3	2·00292 ± 0·00010		A_α	A_σ
					(1)	12·8	1·00
					(2)	0·45	0·19
					(3)	0·18	0·21
NaF			3	2·00023 ± 0·00010	$A_{Na} = 35{\cdot}2 \pm 0{\cdot}3$ $A_F = 8{\cdot}0 \pm 1{\cdot}0$		
NaCl	~10^{17}	290 77	3	1·990 ± 0·010	Not resolved		
KCl	~10^{17}	290 77	3	1·996 ± 0·002	Not resolved		

$I^{233} = 5/2 \quad I^{235} = 7/2$

H.F.S.	Remarks	Ref.
$A^{235} = 176 \pm 1$ $B^{235} = 57{\cdot}5 \pm 1$ $P^{235} = 5{\cdot}5 \pm 0{\cdot}5$ $A^{233} = 378{\cdot}6 \pm 1{\cdot}2$ $B^{233} = 123{\cdot}6 \pm 1{\cdot}0$ $P^{233} = 9{\cdot}9 \pm 1{\cdot}0$	$\mu_N{}^{235}/\mu_N{}^{233} = 0{\cdot}651 \pm 0{\cdot}002$ $Q^{235}/Q^{233} = 1{\cdot}17 \pm 0{\cdot}20$	Dor 57 Hut3 56 Bl3 56
h.f.s. from ^{19}F nuclei		Bl4 56
h.f.s. from ^{19}F nuclei		Bl4 56

$S' = 7/2$

H.F.S.	Remarks	Ref.
	$\mathscr{H} = g\beta\mathbf{H}.\mathbf{S} + B_2{}^0P_2{}^0 + B_4{}^0P_4{}^0 + B_6{}^0P_6{}^0 + B_6{}^6P_6{}^6$	Fi 57
	$\mathscr{H} = g\beta\mathbf{H}.\mathbf{S} + D[S_z{}^2 - \tfrac{1}{3}S(S+1)]$	Abr1 56

$I(^7\text{Li}) = I(^{23}\text{Na}) = I(^{39}\text{K}) = I(^{35,37}\text{Cl}) = 3/2$

Remarks	Ref.
$\Delta H = 76$ gauss. Centres produced by 50 kv x-rays $\Delta H \sim 150$ gauss. Centres produced by neutrons	Je 54 Hut1 49
Centres produced by 180 kv x-rays	Ki 58 Sc1 54
(1), (2), (3) refer to shells of surrounding nuclei, i.e. to ^{7}Li, ^{19}F, ^{7}Li. Double resonance used. Centres produced by neutrons or 50 kv x-rays	Lor 58 Lor 57
h.f.s. from nearest ^{23}Na and ^{19}F nuclei. Centres produced by neutron irradiation	Lor 57
$\Delta H = 160$ gauss. A estimated from line width to be : $A_{Na} = 30$ gauss ; $A_{Cl} = 2{\cdot}6$ gauss Centres produced additively, by x-rays and by electrons	Sc2 51 Kip 53
$\Delta H(^{39}\text{KCl}) = 50$ gauss. A estimated from line width to be : $A_K{}^{39} = 7{\cdot}5$ gauss ; $A_K{}^{41} = 4{\cdot}2$ gauss $\Delta H(^{41}\text{KCl}) = 36$ gauss. A estimated from line width to be : $A_{Cl} = 2{\cdot}6$ gauss Centres produced additively, by x-rays and by electrons	Hut4 52 Sc2 51 Kip 53

F Centres

Formula	No. per cm³	T	λ (cm)	g	H.F.S.
KCl	2×10^{17}	1·2	3		$A_K = 7{\cdot}8$, $B_K = 6{\cdot}9$, $P_K = 0{\cdot}07$ } ^{39}K; $A_{Cl} = 2{\cdot}7$, $B_{Cl} = 2{\cdot}2$ } ^{35}Cl
KI		290	3	$1{\cdot}971 \pm 0{\cdot}001$	
		77			
		4			
KBr	10^{16}–10^{20}	290	3	$1{\cdot}990 \pm 0{\cdot}010$	Not resolved
		77			
MgO	$\sim10^{19}$	290	3	$2{\cdot}0023 \pm 0{\cdot}0001$	$A_{Mg} = 4{\cdot}6$, $B_{Mg} = 3{\cdot}3$

M

Formula	No. per cm³	T	λ (cm)	g	H.F.S.
KCl	$\sim2\times10^{17}$		3	2·000	

Z_1

Formula	No. per cm³	T	λ (cm)	g	H.F.S.
KCl	5×10^{16}		3	$1{\cdot}999 \pm 0{\cdot}005$	

Ag^0

Formula	No. per cm³	T	λ (cm)	g	H.F.S.
KCl		77		$1{\cdot}997 \pm 0{\cdot}003$	$A = 619 \pm 2$, $A_a = 12{\cdot}1 \pm 0{\cdot}2$, $A_e = 0{\cdot}12 \pm 0{\cdot}03$

V Centres $S' = 1/2$ $I(^{19}F) = 1/2$

Formula	No. per cm³	T	λ (cm)	g	H.F.S.
LiF	$\sim10^{17}$	77	3	$g_\parallel = 2{\cdot}0031 \pm 0{\cdot}001$, $g_\perp = 2{\cdot}0230 \pm 0{\cdot}001$	$A_F = 829$, $B_F = 56$ } ^{19}F
NaCl	$\sim10^{17}$	77	3	$g_\parallel = 2{\cdot}0010 \pm 0{\cdot}0001$, $g_\perp = 2{\cdot}045 \pm 0{\cdot}003$	$A_{Cl} = 92$, $B_{Cl} = 8{\cdot}5$ } ^{35}Cl
KCl	$\sim10^{17}$	77	3	$g_\parallel = 2{\cdot}0010 \pm 0{\cdot}0001$, $g_\perp = 2{\cdot}0438 \pm 0{\cdot}001$	$A_{Cl} = 94$, $B_{Cl} = 8{\cdot}5$ } ^{35}Cl
KBr	$\sim10^{17}$	77	3	$g_\parallel = 1{\cdot}980 \pm 0{\cdot}001$, $g_\perp = 2{\cdot}177 \pm 0{\cdot}001$	$A_{Br} = 421$, $B_{Br} = 79$ } ^{81}Br

(*continued*)

Remarks	Ref.
h.f.s. from nearest K and Cl nuclei. Resolved by double-resonance technique	Fe1 57
Centres produced by 1 MeV electrons	
$\Delta H = 200$ gauss. Additively coloured	No 58
$\Delta H = 150$ gauss. A estimated from line width to be : $A_K = 7{\cdot}5$ gauss ; $A_{Br} = 15$ gauss	Tin2 51 Hut4 52 Sc2 51 Kip 53
F centre produced by neutron irradiation or by addition of Mg and x-irradiation. Sextet from single ^{25}Mg, 11 lines from a pair of ^{25}Mg nuclei. $I^{25} = 5/2$ (10% abundance)	Wer2 57

Centre

Remarks	Ref.
Produced by F light on additively coloured crystals. $\Delta H = 96$ gauss	Kaw 58a

Centre

Remarks	Ref.
Crystals doped with Ca, Sr, additively coloured and irradiated with F light. Centre consists of an electron bound to a Ca^{2+} or Sr^{2+} ion. $\Delta H = 79 \pm 8$ gauss	Kaw 58b

Centre

Remarks	Ref.
Crystals doped with Ag and x-irradiated at 77°K. Centre is a Ag atom. A refers to ^{109}Ag ($I = 1/2$). A_s and A_o refer to interaction with Cl nuclei	Ha2 u

$I(^{35,37}\mathrm{Cl}) = I(^{79,81}\mathrm{Br}) = 3/2$

Remarks	Ref.
The g tensor does not possess exactly axial symmetry. $g_\perp$ is the average of g_x and g_y	Sm 56 Cas 58
The axis is the 110 axis	
The results are interpreted as being due to a single ' hole ' associated with a pair of nearest neighbour halide ions, e.g. Cl_2^-	
Centres produced by x-irradiation at 77°K	

V Centres

Formula	No. per cm³	T	λ (cm)	g	H.F.S.
LiF				—	x y z A_3 +1122 −212 −268 A_1A_3 + 378 +142 + 80 (in gauss)
H centre in KCl		20		$g_1 = 2{\cdot}0224 \pm 0{\cdot}0001$ $g_\perp = 2{\cdot}0023 \pm 0{\cdot}0005$	$A_{Cl}^1 = 103$ $B_{Cl}^1 = 0$ } ^{35}Cl $A_{Cl}^2 = 6{\cdot}9$ $B_{Cl}^2 = 2{\cdot}7$ } ^{35}Cl
KCl		77		$g_1 = 2{\cdot}005$ $g_\perp = 2{\cdot}04$	$A_{Cl} = 88{\cdot}5$ $B_{Cl} \sim 16$
KBr		77		$g_1 = 1{\cdot}993$	$A_{Cl} = 415$
KCl		208		$g_1 = 2{\cdot}196 \pm 0{\cdot}004$ $g_\perp = 2{\cdot}038 \pm 0{\cdot}004$	$A = 41 \pm 1$ $B = 30 \pm 4$ $A_{Cl} < 4$ $B_{Cl} = 30 \pm 4$

Damage

Formula	No. per cm³	T	λ (cm)	g	H.F.S.
Diamond irradiated with neutrons or electrons		290		(*a*) $g = 2{\cdot}0028 \pm 0{\cdot}004$ isotropic	No h.f.s.
		90	3		
		20	1·2	(*b*) $g_1 = 2{\cdot}0024 \pm 0{\cdot}0003$ $g_\perp = 2{\cdot}0031 \pm 0{\cdot}0003$	No h.f.s.
Quartz irradiated with neutrons, γ-rays, X-rays, electrons		90	3	$g_1 = 2{\cdot}060 \pm 0{\cdot}005$	$A = 4{\cdot}8$ } h.f.s. due to $B = 5{\cdot}6$ } ^{27}Al $P = -0{\cdot}4$ impurity
		20	1·2	$g_\perp = 2{\cdot}000 \pm 0{\cdot}005$	
Quartz, neutron irradiated, powder or glass		290		(*a*) $g = 2{\cdot}0013 \pm 0{\cdot}0006$ (*b*) $g = 2{\cdot}0090 \pm 0{\cdot}0007$	
$NaNO_3$ X-irradiated		290		$2{\cdot}000 \pm 0{\cdot}003$	Triplet due to ^{14}N
		77			$A = 63{\cdot}0 \pm 2{\cdot}5$
$La_2Mg_3(NO_3)_{12}.24H_2O$ γ-irradiated		290		$g_x = 2{\cdot}005 \pm 0{\cdot}002$	$A_x = 63 \pm 3$ } From ^{14}N
		90		$g_y = 2{\cdot}005 \pm 0{\cdot}002$	$A_y = 57 \pm 3$ } From ^{14}N
		20		$g_z = 1{\cdot}993 \pm 0{\cdot}002$ (averaged results)	$A_z = 48 \pm 4$ (averaged results)

Impurities in

Formula	No. of centres	T	λ	g	Fine structure
Li in Si	7×10^{16} per cm³	20	3 cm	1·999	
		4	300 Mc/s		

(*continued*)

Remarks	Ref.
This centre is F_3^{2-} and is produced by x-irradiation at 77°K and subsequent warming to 130°K. Associated with V_4 centre. The A_{123} are the contributions to the h.f.s. from each ^{19}F nucleus	Co 57
This centre is similar to the Cl_3^- centre above, but the electron interacts with two further ^{35}Cl nuclei at the ends of a linear chain of four. A^1 and B^1 give the contribution to the h.f.s. from the two-centre ^{35}Cl nuclei and A^2, B^2 that from the outer pair. The axis is the 110 axis. Centres produced by x-irradiation at 20°K	Ka 58
Crystals doped with Ba, Sr, Ca and x-irradiated at 77°K. Centre is Cl_3^- with 100 axis	Ni u
Centre is Br_3^- with 100 axis. Produced as above	Ni u
Crystals doped with Ag, x-irradiated at 77°K and warmed to 208°K. Centre is a hole associated with a Ag^+ ion and two Cl^- ions. A and B refer to ^{109}Ag ($I = 1/2$)	Ha2 u

Centres

Remarks	Ref.
Possibly due to single vacancies or interstitial carbon atoms	Gr 54
$\mathscr{H} = g\beta\mathbf{H}.\mathbf{S} + DS_z^2$, $S = 1$, $D = 100 \pm 2$. Possibly due to C_3 molecules	
Spectrum interpreted as due to an electron ($S' = 1/2$) spending most of its time on an oxygen atom but a small part on an Al impurity atom in a Si lattice site. $I^{27} = 5/2$	Gr 54
Two lines in powder or glass. In single crystals there are two sets of lines which can be associated with these	We 56a We 56b
Attributed to NO_2 or NO_3 molecules. Other lines present but not interpreted	Ard 55
Resonance attributed to NO_2 molecules, $S' = 1/2$, $I^{14} = 1$. The centres were produced by ^{60}Co γ-rays or by growing crystals containing radioactive nuclei	Bl2 57

Si $S' = 1/2$

H.F.S.	Remarks	Ref.
Not resolved	$\Delta H = 1{\cdot}5$ gauss, explained by h.f.s. from ^{29}Si assuming electron orbit encloses ~600 Si atoms	Ho3 54

Impurities in Si

Formula	No. of centres	T	λ cm	g	Fine structure
Li in Si		1·2	3		
^{31}P and ^{32}P in Si	^{31}P $\sim 10^{17}$ ^{32}P 4×10^{14}	4 1·2	3 1·2	$2{\cdot}0004 \pm 0{\cdot}0005$	
Sb in Si	4×10^{17} per cm^3	4	1·2	$1{\cdot}99853 \pm 0{\cdot}00001$	
B in Si		4	3	(*a*) $2{\cdot}004 \pm 0{\cdot}001$ (*b*) $2{\cdot}002 \pm 0{\cdot}001$	
As in Si	$\sim 10^{17}$	4	3 1·2	$2{\cdot}0004 \pm 0{\cdot}0005$	
P, As in Si		1·2	3	$\bar{g} = 2{\cdot}0053 \pm 0{\cdot}0001$ (mean g value)	

Conduction

Metal	T	λ	g	ΔH	
Lithium	4–290	3 cm, 400 Mc/s to 25 Mc/s	$2{\cdot}0023 \pm 0{\cdot}0001$	Bulk metal, 4–290°K 0·3 gauss (Fe5 55)	Frozen Li–ammonia solution, 4°K, 0·4 gauss 77°K, sharp rise from 1·5 gauss to 4·5 gauss 77–150°K, 4·5 gauss (Lev 56, Fe4 54)
Sodium	4–385	3 cm, 400 Mc/s to 25 Mc/s	$2{\cdot}0015 \pm 0{\cdot}0002$	385°K, 9·7 gauss 290°K, 10 gauss 200°K, 5 gauss 77°K, 2 gauss 4°K, 0·1 gauss	
Potassium	4–220	300 Mc/s	1·99	150°K, 20 gauss 4°K, 5 gauss	
Cesium	4–30	300 Mc/s	$1{\cdot}93 \pm 0{\cdot}02$	4–30°K, 13 gauss	
Beryllium	4–290	3 cm, 300 Mc/s	$2{\cdot}0032 \pm 0{\cdot}0001$	4–290°K, 20 gauss	
Silicon (doped with P)	77	3 cm, 200 Mc/s	~ 2	$\Delta H = 8$ gauss	

(*continued*)

H.F.S.	Remarks	Ref.
$A^{7} = 0{\cdot}28$	h.f.s. resolved by double resonance. There is also structure from ^{29}Si nuclei. $I^{7} = 3/2$	Fe1 56
$A^{31} = 39$ $A^{33} = 4{\cdot}380 \pm 0{\cdot}003$	The value of A^{33} was obtained by the double-resonance method. Resonance obtained from 2, 3, 4 coupled spins $I^{31} = 1/2$, $I^{33} = 1$	Fe2 55 Fl1 54 Fl2 54 Fe3 57
$A^{121} = 62{\cdot}267 \pm 0{\cdot}005$ $A^{123} = 33{\cdot}839 \pm 0{\cdot}004$	I^{121} (57·2%) = 5/2 I^{123} (42·7%) = 7/2 } Double resonance (Ei 58)	Fl1 54 Fl2 54 Ei 58
	(*a*) Neutron irradiated ΔH = 12 gauss (*b*) Partially annealed ΔH = 3 gauss	Sch 55
$A^{75} = 69 \pm 2$	Compression of sample increases intensity of spectrum. So does addition of boron. $I^{75} = 3/2$	Fl1 54 Fl2 54 Ho1 54 Ho2 56
$A(^{29}\text{Si}) \simeq 140$	Irradiated with 0·5 Mev electrons. Attributed to Si atoms displaced along 100 directions. I^{29} (4·7%) = 1/2	Bem 58

Electrons

Remarks	Ref.
Samples used: Particles dispersed in oil, paraffin and precipitated from Li–ammonia solution. Also bulk metal. (Ref. Lev 56 gives results on molten metal)	Lev 56 Fe5 55 Ga1 57 Fe6 55 Fe4 54 Gu 54
Samples as for Li. Where authors give different line widths, the minimum values are quoted	Lev 56 Fe5 55 Ga1 57 Fe6 55 So 54 Fe4 54 Gu 54
Frozen K–ammonia solution	Lev 56 Fe6 55
Line disappears above 30°K. Samples: frozen Cs–ammonia solution	Lev 56 Fe6 55
Particles in paraffin	Fe5 55 Fe5 54
Particles in paraffin	Abg 58

ADDENDUM

$3d^2\ V^{3+}$ $S' = 1$

Formula	Dilution	T	λ (cm)	g	Fine structure
Al_2O_3		4·2	2–0·8	$g_\parallel = 1{\cdot}92 \pm 0{\cdot}01$	$D \sim 10\ cm^{-1}$

$3d^3\ Cr^{3+}$ $S' = 3/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
MgO	$\sim 1 : 10^5$	290 90 20	3 1·2	1·98	$D = 310 \pm 20$ $E = 2200 \pm 100$

$3d^5\ Mn^{2+}$ $S' = 5/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
ZnO	$\sim 1 : 10^3$	77	1·2	$2{\cdot}0016 \pm 0{\cdot}0006$	$D = -216{\cdot}9 \pm 2{\cdot}2$ $F = -6{\cdot}0 \pm 1{\cdot}5$
CdS	$1 : 10^3$	290	1·2	$2{\cdot}0029 \pm 0{\cdot}0006$	$D = +8{\cdot}2 \pm 2{\cdot}2$ $F = -4{\cdot}2 \pm 1{\cdot}5$
NaCl	$\sim 1 : 10^4$	290	3	$2{\cdot}0012 \pm 0{\cdot}0008$	$D = +128{\cdot}2 \pm 1$ $E = +47{\cdot}8 \pm 0{\cdot}4$

$3d^5\ Fe^{3+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
$BaTiO_3$	$\sim 1 : 10^4$	160°c 290	3		$a = 94$ $a \simeq \mp 170$* $D \simeq \pm 900$*
Al_2O_3	$1 : 10^3$	290	1·2 to 0·8	$2{\cdot}003 \pm 0{\cdot}001$	$D = 1685 \pm 3$ $\lvert a \rvert = 262 \pm 2$ $\lvert a-F \rvert = 334 \pm 2$

$3d^6\ Fe^{2+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
Silicon		10	2	2·0699	

$3d^7\ Co^{2+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$Bi_2Mg_3(NO_3)_{12}.24H_2O$	$5 : 10^3$	4	3	(1) $g_\parallel = 4{\cdot}145 \pm 0{\cdot}002$ $g_\perp = 4{\cdot}415 \pm 0{\cdot}002$ (2) $g_\parallel = 7{\cdot}20 \pm 0{\cdot}01$ $g_\perp = 2{\cdot}39 \pm 0{\cdot}002$	

* Figures calculated from ground

$I^{51}(99{\cdot}8\%) = 7/2$

H.F.S.	Remarks	Ref.
$A = 193 \pm 3$	$\Delta M = 2$ transitions observed	16

$I^{53}(9{\cdot}5\%) = 3/2$

H.F.S.	Remarks	Ref.
	Attributed to Cr^{3+} with positive ion vacancy in n.n. positive ion site. z axis 100 type $\perp$ to line joining	4

$I^{55}(100\%) = 5/2$

H.F.S.	Remarks	Ref.
$A = -76{\cdot}0 \pm 0{\cdot}4$	For spin Hamiltonian see Mn^{2+} in ZnS (hexagonal)	2
$A = -65{\cdot}3 \pm 0{\cdot}4$	Also reports structure from n.n. ^{111}Cd and ^{113}Cd nuclei. For spin Hamiltonian see Mn^{2+} in ZnS (hexagonal)	2
$A = -82{\cdot}6 \pm 0{\cdot}8$	Attributed to Mn^{2+} with positive ion vacancy in n.n. positive ion site. z axis line joining, i.e. 110 type. Crystals from aqueous solution quenched from high temperature	12

$S' = 5/2$

H.F.S.	Remarks	Ref.
	Attributed to Fe^{3+} in Ti^{3+} sites. Cubic phase 160°c. Tetragonal phase 290°K	5
	Two sets of lines having same D axis but different cubic axes	8

H.F.S.	Remarks	Ref.
$A^{57} = 7{\cdot}0$	Enriched ^{57}Fe I = $\frac{1}{2}$	10

$I^{59}(100\%) = 7/2$

H.F.S.	Remarks	Ref.
(1) $A = 95 \pm 3$ $B = 103 \pm 3$ (2) $A = 302 \pm 6$ $B < 1$	(1) and (2) are inequivalent ions. Results also given of variation of A, B with concentration of added $Ce_2Mg_3(NO_3)_{12} \cdot 24H_2O$	3

state splittings quoted in paper.

$3d^7\ Co^{2+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
$La_2Mg_3(NO_3)_{12}.24H_2O$	$5:10^3$	4	3	(1) $g_\parallel = 4{\cdot}050 \pm 0{\cdot}002$ $g_\perp = 4{\cdot}430 \pm 0{\cdot}002$ (2) $g_\parallel = 7{\cdot}23 \pm 0{\cdot}01$ $g_\perp = 2{\cdot}310 \pm 0{\cdot}002$	

$3d^9\ Cu^{2+} \quad S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
AgCl (chlorinated)		290 77	3	(*a*) $g_\parallel = 2{\cdot}00 \pm 0{\cdot}02$ $g_\perp = 2{\cdot}28 \pm 0{\cdot}02$ (*b*) $g_\parallel = 2{\cdot}26 \pm 0{\cdot}02$ $g_\perp = 2{\cdot}07 \pm 0{\cdot}02$ (*c*) $g_\parallel = 2{\cdot}30 \pm 0{\cdot}02$ $g_\perp = 2{\cdot}07 \pm 0{\cdot}02$	

$4f^1\ Ce^{3+}$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	2 : 100	4	3 1	$g_\parallel = 4{\cdot}0366 \pm 0{\cdot}0015$ $g_\perp = 0{\cdot}17 \pm 0{\cdot}08$	

$4f^2\ Pr^{3+} \quad S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	2 : 100	4	3 1	$g_\parallel = 1{\cdot}035 \pm 0{\cdot}005$ $g_\perp = 0{\cdot}1 \pm 0{\cdot}15$	$\Delta \simeq 200$

$4f^3\ Nd^{3+} \quad S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	$2:10^3$	4	3 1	$g_\parallel = 3{\cdot}996 \pm 0{\cdot}001$ $g_\perp = 1{\cdot}763 \pm 0{\cdot}001$	
$La(NO_3)_3.6H_2O$	1 : 200	13–20	3	$g_x = 3{\cdot}88 \pm 0{\cdot}01$ $g_y = 1{\cdot}72 \pm 0{\cdot}01$ $g_z = 0{\cdot}74 \pm 0{\cdot}01$	

(*continued*)

H.F.S.	Remarks	Ref.
(1) $A = 80{\cdot}7 \pm 3$ $B = 103{\cdot}3 \pm 3$ (2) $A = 278{\cdot}8 \pm 6$ $B \leqslant 1$		3

$I^{63}(69{\cdot}1\%) = 3/2 \quad I^{65}(30{\cdot}9\%) = 3/2$

H.F.S.	Remarks	Ref.
(*a*) $A = 10 \pm 10$ $B = 91 \pm 5$ $P = 19 \pm 5$	(*a*) Attributed to Cu^{2+} with vacancy in n.n. positive ion site. z axis 110	14
(*b*) $A = 101 \pm 5$ $B = 43 \pm 5$ $P = 20 \pm 5$	(*b*) Cu^{2+} with vacancy in n.n.n. positive ion site. z axis 100	
(*c*) $A = 113 \pm 5$ $B = 44 \pm 5$ $P = 20 \pm 5$	(*c*) z axis 100. Axial field attributed to Jahn–Teller effect. A, B, P mean values for the two nuclei	

$S' = 1/2$

H.F.S.	Remarks	Ref.
		6

$I^{141}(100\%) = 5/2$

H.F.S.	Remarks	Ref.
$A = 502 \pm 3$	$\mathscr{H} = \beta \mathbf{H}.g.\mathbf{S} + AS_zI_z + B(S_xI_x + S_yI_y) + \Delta_xS_x + \Delta_yS_y$ $\Delta = (\Delta_x^2 + \Delta_y^2)^{1/2}$	6

$I^{143}(12{\cdot}2\%) = 7/2 \quad I^{145}(8{\cdot}3\%) = 7/2$

H.F.S.	Remarks	Ref.
$A^{143} = 425 \pm 2$ $B^{143} = 167 \pm 1$ $P^{143} < 1$ $A^{145} = 264 \pm 2$ $B^{145} = 104 \pm 1$ $P^{145} < 1$		6
$A_x^{143} = 432 \pm 2$ $A_y^{143} = 193 \pm 2$ $A_z^{143} = 82 \pm 10$ $A_x^{145} = 270 \pm 2$ $A_y^{145} = 119 \pm 2$ $A_z^{145} = 51 \pm 10$		13

$4f^5\ Sm^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	2 : 100	4	3	$g_\parallel = 0{\cdot}5841 \pm 0{\cdot}0003$	
			1	$g_\perp = 0{\cdot}6127 \pm 0{\cdot}0006$	

$4f^7\ Gd^{3+}$ $S' = 7/2$

Formula	Dilution	T	λ (cm)	g	Fine structure	
$LaCl_3$	$2 : 10^3$	4	3	$1{\cdot}991 \pm 0{\cdot}001$		
			1			
$LaCl_3.7H_2O$		290	3	$g_x = 1{\cdot}998 \pm 0{\cdot}003$	290°K	80°K
		80		$g_y = 2{\cdot}000 \pm 0{\cdot}003$	$b_2^0 = 131{\cdot}3 \pm 0{\cdot}5$	99 ± 30
				$g_z = 1{\cdot}989 \pm 0{\cdot}003$	$b_2^2 = 75{\cdot}2 \pm 2$	115 ± 7
					$b_4^0 = 2 \pm 1$	
SrS	$1 : 10^4$		3			
ThO_2	$< 1 : 10^4$	290	3	$1{\cdot}9913 \pm 0{\cdot}0005$	$b_4^0 = 54{\cdot}7 \pm 0{\cdot}1$	
					$b_6^0 = 0{\cdot}26 \pm 0{\cdot}08$	
		90		$1{\cdot}991 \pm 0{\cdot}001$	$b_4^0 = 56{\cdot}0 \pm 0{\cdot}2$	
					$b_6^0 = 0{\cdot}4 \pm 0{\cdot}2$	

$4f^8\ Tb^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	$5 : 10^3$	4	3	$g_\parallel = 17{\cdot}78 \pm 0{\cdot}01$	$\Delta = 2010$
	2 : 100		1	$g_\perp < 0{\cdot}1$	

$4f^{10}\ Ho^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	$5 : 10^3$	4	3	$g_\parallel = 16{\cdot}01 \pm 0{\cdot}018$	
			1	$g_\perp \sim 0$	

$4f^{11}\ Er^{3+}$ $S' = 1/2$

Formula	Dilution	T	λ (cm)	g	Fine structure
$LaCl_3$	$5 : 10^3$	4	3	$g_\parallel = 1{\cdot}989 \pm 0{\cdot}001$	
			1	$g_\perp = 8{\cdot}757 \pm 0{\cdot}002$	

V centres $S' = 1/2$

Formula	No. per cm^3	T	λ (cm)	g
LiF		290	3	

$I^{147}(15{\cdot}1\%) = 7/2$ $I^{149}(13{\cdot}8\%) = 7/2$

H.F.S.	Remarks	Ref.
$A^{147} = 60{\cdot}7 \pm 0{\cdot}2$ $B^{147} = 245 \pm 1$ $A^{149} = 49{\cdot}9 \pm 0{\cdot}2$ $B^{149} = 202 \pm 1$		6

$I^{155}(14{\cdot}7\%) = 3/2$ $I^{157}(15{\cdot}2\%) = 3/2$

H.F.S.	Remarks	Ref.
$A^{155} = 3{\cdot}8 \pm 0{\cdot}2$ $A^{157} = 5{\cdot}0 \pm 0{\cdot}2$		6
		15
$A^{155} = 3{\cdot}68 \pm 0{\cdot}06$ gauss $A^{157} = 5{\cdot}00 \pm 0{\cdot}05$ gauss	$\mu_N{}^{155}/\mu_N{}^{157} = 0{\cdot}73 \pm 0{\cdot}03$	11
$A^{155} = 3{\cdot}93 \pm 0{\cdot}15$ $A^{157} = 5{\cdot}30 \pm 0{\cdot}12$	Cubic field $\mu_N{}^{155}/\mu_N{}^{157} = 0{\cdot}744 \pm 0{\cdot}007$	9

$I^{159}(100\%) = 3/2$

H.F.S.	Remarks	Ref.
$A = 2120 \pm 30$	$\mathscr{H} = \beta \mathbf{H}.g.\mathbf{S} + AS_zI_z + B(S_xI_x + S_yI_y) + \Delta_xS_x + \Delta_yS_y$ $\Delta = (\Delta_x{}^2 + \Delta_y{}^2)^{1/2}$	6

$I^{165}(100\%) = 7/2$

H.F.S.	Remarks	Ref.
$A = 3510 \pm 70$		6

$I^{187}(22{\cdot}9\%) = 7/2$

H.F.S.	Remarks	Ref.
$A = 66{\cdot}4 \pm 0{\cdot}3$ $P = 8{\cdot}6$		6

$I(^{19}\mathrm{F}) = 1/2$

H.F.S.	Remarks	Ref.
$A_F = 951$ gauss	Centres produced by neutron irradiation, attributed to $F_3{}^-$. cf. $A_F = 887$ gauss for $F_2{}^-$ produced by x-rays (Cas 58)	7

References (to main body of tables)

Ab1 Abe, H., 1957, *J. Phys. Soc. Japan*, **12**, 435.

Ab2 Abe, H., Nagano, H., Nagusa, M., and Oshima, K., 1956, *J. Chem. Phys.*, **25**, 378.

Abg Abragam, A., Combrisson, J., and Solomon, I., 1958, *C.R. Acad. Sci., Paris*, **246**, 1035.

Abr1 Abraham, M., Cunningham, B. B., Jeffries, C. D., and Kedzie, R. W., 1956, *Bull. Amer. Phys. Soc.*, **1**, 396.

Abr2 Abraham, M., Jeffries, C. D., Kedzie, R. W., and Wallman, J. C., 1957, *Phys. Rev.*, **106**, 1357.

Abr3 Abraham, M., Kedzie, R. W., and Jeffries, C. D., 1957, *Phys. Rev.*, **108**, 58.

Ard Ard, W. B., 1955, *J. Chem. Phys.*, **23**, 1967.
Bagguley, D. M. S., and Owen, J., 1957, *Rep. Progr. Phys.*, **20**, 304 (London : Physical Society).

Ba1 Baker, J. M., 1956, *Proc. Phys. Soc.* B, **69**, 633.

Ba2 Baker, J. M., and Bleaney, B., 1955 a, *Proc. Phys. Soc.* A, **68**, 936 ; 1955 b, *Ibid.*, **68**, 1090.

Ba3 Baker, J. M., Bleaney, B., and Bowers, K. D., 1956, *Proc. Phys. Soc.* B, **69**, 1205.

Ba4 Baker, J. M., Bleaney, B., and Hayes, W., 1958, *Proc. Roy. Soc.* A, **247**, 141.

Ba5 Baker, J. M., Bleaney, B., Llewellyn, P. M., and Shaw, P. F. D., 1956, *Proc. Phys. Soc.* A, **69**, 353.

Bem Bemski, G., Feher, G., and Gere, E., 1958, *Bull. Amer. Phys. Soc.*, **3**, 135.

Be1 Bennett, J. E., Gibson, J. F., and Ingram, D. J. E., 1957, *Proc. Roy. Soc.* A, **240**, 67.

Be2 Bennett, J. E., and Ingram, D. J. E., 1955, *Nature, Lond.*, **175**, 130 ; 1956, *Phil. Mag.*, **1**, 970.

Bl1 Bleaney, B., and Hayes, W., 1957, *Proc. Phys. Soc.* B, **70**, 626.

Bl2 Bleaney, B., Hayes, W., and Llewellyn, P. M., 1957, *Nature, Lond.*, **179**, 140.

Bl3 Bleaney, B., Hutchison, C. A., Llewellyn, P. M., and Pope, D. F. D., 1956, *Proc. Phys. Soc.* B, **69**, 1167.

Bl4 Bleaney, B., Llewellyn, P. M., and Jones, D. A., 1956, *Proc. Phys. Soc.* B, **69**, 858.

I Bleaney, B., and Stevens, K. W. H., 1953, *Rep. Progr. Phys.*, **16**, 108 (London : Physical Society).

II Bowers, K. D., and Owen, J., 1955, *Rep. Progr. Phys.*, **18**, 304 (London : Physical Society).

Ca Carlson, E. H., and Spence, R. D., 1956, *J. Chem. Phys.*, **24**, 471.

Cas Castner, T. G., and Kanzig, W., 1958, *J. Phys. Chem. Solids*, **3**, 178.

Co Cohen, M. H., Kanzig, W., and Woodruff, T. O., 1957, *Phys. Rev.*, **108**, 1096.

Coo Cooke, A. H., and Park, J. G., 1956, *Proc. Phys. Soc.* A, **69**, 282.

Da Davis, C. F., and Strandberg, M. W. P., 1957, *Phys. Rev.*, **105**, 447.

Do1 Dobrov, W., and Jeffries, C. D., 1957, *Phys. Rev.*, **108**, 60.

Do2 Dobrowolski, W., Jones, R. V., and Jeffries, C. D., 1956 a, *Phys. Rev.*, **101**, 1001 ; 1956 b, *Ibid.*, **104**, 1378.

Dor Dorain, P. B., Hutchison, C. A., and Wong, E., 1957, *Phys. Rev.*, **105**, 1307.

Ei Eisinger, J., and Feher, G., 1958, *Phys. Rev.*, **109**, 1172.
Elliott, R. J., and Stevens, K. W. H., 1953, *Proc. Roy. Soc.* A, **219**, 387.

Fe1 Feher, G., 1956, *Phys. Rev.*, **103**, 834 ; 1957, *Ibid.*, **105**, 1122.

Fe2 Feher, G., Fletcher, R. C., and Gere, A. E., 1955, *Phys. Rev.*, **100**, 1784.

Fe3 Feher, G., Fuller, C. S., and Gere, A. E., 1957, *Phys. Rev.*, **107**, 1462.

Fe4 Feher, G., and Griswold, T. W., 1954, *Phys. Rev.*, **93**, 952.

Fe5 Feher, G., and Kip, A. F., 1954, *Phys. Rev.*, **95**, 1343 ; 1955, *Ibid.*, **98**, 337.

Fe6 Feher, G., and Levy, R. A., 1955, *Phys. Rev.*, **98**, 264.

Fi Fields, P., Friedman, A., and Smaller, B., 1957, *Phys. Rev.*, **105**, 757.

Fl1 Fletcher, R. C., Yager, W. A., Pearson, G. L., Holden, A. N., Read, W. T., and Merritt, F. R., 1954, *Phys. Rev.*, **94**, 1392.

Fl2 Fletcher, R. C., Yager, W. A., Pearson, G. I., and Merritt, F. R., 1954, *Phys. Rev.*, **95**, 844.

Fo FORRESTER, P. A., and SCHNEIDER, E. E., 1956, *Proc. Phys. Soc.* B, **69**, 833.

Fu FUKUDA, K., UCHIDA, Y., and YOSHIMURA, H., 1958, *J. Phys. Soc., Japan*, **13**, 971.

Ga1 GARIF'IANOV, N. S., 1957, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **32**, 149 (see *Soviet Physics, J.E.T.P.*, 1957, **5**, 111).

Ga2 GARIF'IANOV, N. S., and ZARIPOV, M. M., 1955, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **28**, 629 (see *Soviet Physics, J.E.T.P.*, 1955, **1**, 567).

Ge GERRITSEN, H. J., OKKES, R., BOLGER, B., and GORTER, C. J., 1955, *Physica*, **21**, 629.

Geu GEUSIC, J. E., 1956, *Phys. Rev.*, **102**, 1252.

Gr GRIFFITHS, J. H. E., OWEN, J., and WARD, I. M., 1955, *Report of the Conference on Defects in Crystalline Solids*, 1954, p. 81 (London : Physical Society).

Gu GUTOWSKY, H. S., and FRANK, P. J., 1954, *Phys. Rev.*, **94**, 1067.

Ha HAYES, W., and JONES, D. A., 1958, *Proc. Phys. Soc.*, **71**, 503.

Hal HALFORD, D., HUTCHISON, C. A., and LLEWELLYN, P. M., 1958, *Phys. Rev.*, **110**, 284.

Ho1 HONIG, A., 1954, *Phys. Rev.*, **96**, 234.

Ho2 HONIG, A., and COMBRISSON, J., 1956, *Phys. Rev.*, **102**, 917.

Ho3 HONIG, A., and KIP, A. F., 1954, *Phys. Rev.*, **95**, 1686.

Hor HORNIG, A. W., JAYNES, E. T., and WEAVER, H. E., 1954, *Phys. Rev.*, **96**, 1703.

Hut1 HUTCHISON, C. A., 1949, *Phys. Rev.*, **75**, 1769.

Hut2 HUTCHISON, C. A., JUDD, B. R., and POPE, D. F. D., 1957, *Proc. Phys. Soc.* B, **70**, 514.

Hut3 HUTCHISON, C. A., LLEWELLYN, P. M., WONG, E., and DORAIN, P., 1956, *Phys. Rev.*, **102**, 292.

Hut4 HUTCHISON, C. A., and NOBLE, G. A., 1952, *Phys. Rev.*, **87**, 1125.

In INGRAM, D. J. E., and BENNETT, J. E., 1954, *J. Chem. Phys.*, **22**, 1136.

Ja JARRETT, H. S., 1958, *J. Chem. Phys.*, **28**, 1260.

Je JEN, C. K., and LORD, N. W., 1954, *Phys. Rev.*, **96**, 1150.

Jo JONES, R. V., DOBROWOLSKI, W., and JEFFRIES, C. D., 1956, *Phys. Rev.*, **102**, 738.

Ka KANZIG, W., and WOODRUFF, T. O., 1958, *Phys. Rev.*, **109**, 220.

Kaw KAWAMURA, H., and ISHIWATARI, K., 1958 a, *J. Phys. Soc. Japan*, **13**, 33 ; 1958 b, *Ibid.*, **13**, 574.

Ke KEDZIE, R. W., ABRAHAM, M., and JEFFRIES, C. D., 1957, *Phys. Rev.*, **108**, 54.

Kel KELLER, S. P., GELLES, I. L., and SMITH, W. V., 1958, *Phys. Rev.*, **110**, 850.

Ki KIM, Y. W., KAPLAN, R., and BRAY, P. J., 1958, *Bull. Amer. Phys. Soc.*, **3**, 178.

Kip KIP, A. F., KITTEL, C., LEVY, R. A., and PORTIS, A. M., 1953, *Phys. Rev.*, **91**, 1066.

Le LEECH, J. W., and MANUEL, A. J., 1956, *Proc. Phys. Soc.* B, **69**, 210.

Lev LEVY, R. A., 1956, *Phys. Rev.*, **102**, 31.

Lor LORD, N. W., 1957, *Phys. Rev.*, **105**, 756 ; 1958, *Phys. Rev. Let.*, **1**, 170.

Lo1 LOW, W., 1955, *Phys. Rev.*, **98**, 426 ; 1956 a, *Ibid.*, **101**, 1827 ; 1956 b, *Ibid.*, **103**, 1309 ; 1956 c, *Proc. Phys. Soc.* B, **69**, 837 ; 1956 d, *Ibid.*, **69**, 1169 ; 1957 a, *Phys. Rev.*, **105**, 793 ; 1957 b, *Ibid.*, **105**, 801 ; 1958 a, *Ibid.*, **109**, 247 ; 1958 b, *Ibid.*, **109**, 256 ; 1958 c, *Ibid.*, **109**, 265.

Lo2 LOW, W., and LLEWELLYN, P. M., 1958, *Phys. Rev.*, **110**, 842.

Lo3 LOW, W., and SHALTIEL, D., 1958, *Phys. Rev. Let.*, **1**, 51.

Ma1 MANENKOV, A. A., and PROKHOROV, A. M., 1955, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **28**, 762 (see *Soviet Physics, J.E.T.P.*, 1955, **1**, 611) ; 1956 a, *Dok. Akad. Nauk.* (*U.S.S.R.*), **107**, 402 ; 1956 b, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **31**, 346 (see *Soviet Physics, J.E.T.P.*, 1957, **4**, 288).

Ma2 MANENKOV, A. A., PROKHOROV, A. M., TRUKHLYAEV, P. S., and YAKOVLEV, G. N., 1957, *Dok. Akad. Nauk.* (*U.S.S.R.*), **112**, 623.

Mat MATARESE, L. M., and KIKUCHI, C., 1955, *Phys. Rev.*, **100**, 1243 ; 1956, *J. Chem. Phys. Solids*, **1**, 117.

No NOBLE, G. A., 1958, *Bull. Amer. Phys. Soc.*, **3**, 178.

Ok OKAMURA, T., and DATE, M., 1954, *Phys. Rev.*, **94**, 314.

Os OSHIMA, K., ABE, H., NAGANO, H., and NAGUSA, M., 1955, *J. Chem. Phys.*, **23**, 1721.

Ow OWEN, J., BROWNE, M. E., ARP, V., and KIP, A. F., 1957, *J. Chem. Phys. Solids*, **2**, 85.

Pa PARK, J. G., 1958, *Proc. Roy. Soc.* A, **245**, 118.

Ry RYTER, C., and LACROIX, R., 1956, *C.R. Acad. Sci., Paris*, **242**, 2812.

Sc1 Schneider, E. E., 1954, *Phys. Rev.*, **93**, 919.
Sc2 Schneider, E. E., and England, T. S., 1951, *Physica*, **17**, 221.
Sch Schulz-Du Bois, E., Nisenoff, M., Fan, H. Y., and Lark-Horovitz, K., 1955, *Phys. Rev.*, **98**, 1561.
Si Singer, L. S., 1955, *J. Chem. Phys.*, **23**, 379.
Sm Smaller, B., Delbecq, C., and Yuster, P., 1956, *Bull. Amer. Phys. Soc.*, **1**, 324.
So Solt, I. H., and Strandberg, M. W. P., 1954, *Phys. Rev.*, **95**, 607.
Ti Ting, Y., Farringer, L. D., and Williams, D., 1955, *Phys. Rev.*, **97**, 1037.
Tin1 Tinkham, M., 1956, *Proc. Roy. Soc.* A, **236**, 535.
Tin2 Tinkham, M., and Kip, A. F., 1951, *Phys. Rev.*, **83**, 657.
Va Van Wieringen, J. S., 1955, *Disc. Faraday Soc.*, **19**, 118.
Wa1 Walsh, W. M., 1958, *Bull. Amer. Phys. Soc.*, **3**, 178.
Wa2 Walsh, W. M., and Bloembergen, N., 1957, *Phys. Rev.*, **107**, 904.
Wat Watkins, G. D., 1957, *Bull. Amer. Phys. Soc.*, **2**, 345; 1958, *Phys. Rev.*, **110**, 986.
We Weeks, R. A., 1956 a, *Bull. Amer. Phys. Soc.*, **1**, 12; 1956 b, *J. Appl. Phys.*, **27**, 1376.
Wer1 Wertz, J. E., and Auzins, P., 1957, *Phys. Rev.*, **106**, 484.
Wer2 Wertz, J. E., Auzins, P., Weeks, R. A., and Silsbee, R. H., 1957, *Phys. Rev.*, **107**, 1535.
Wo Woodbury, H. H., and Ludwig, G. W., 1958, *Phys. Rev. Let.*, **1**, 16.
Za Zaripov, M. M., and Shamonin, I. I., 1956, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **30**, 291 (see *Soviet Physics, J.E.T.P.*, 1956, **3**, 171).

Unpublished Results

Ba1 u Baker, J. M.
Ba2 u Baker, J. M.
Ha1 u Hayes, W.
Ha2 u Hayes, W., Delbecq, C. J., and Yuster, D. H.
Ni u Nichols, G. M., and Hayes, W.
Or u Orton, J. W.
Pa u Partridge, M. F.

References (to Addendum)

1 Abraham, M., Jeffries, C. D., Kedzie, R. W., and Wallman, J. C., 1958, *Phys. Rev.*, **112**, 725.
2 Dorain, P. B., 1958, *Phys. Rev.*, **112**, 1058.
3 Gager, W. B., Jasram, P. S., and Daunt, J. G., 1958, *Phys. Rev.*, **111**, 803.
4 Griffiths, J. H. E., and Orton, J. W., 1959, *Proc. Phys. Soc.*, **73**, in the press.
5 Hornig, A. W., Rempel, R. C., and Weaver, H. E., 1958, *Phys. Rev. Let.*, **1**, 284.
6 Hutchison, C. A., and Wong, E., 1958, *J. Chem. Phys.*, **29**, 754.
7 Kim, Y. W., Kaplan, R., and Bray, P. J., 1958, *Phys. Rev.*, **111**, 1468.
8 Kornienko, L. S., and Prokhorov, A. M., 1957, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **33**, 805. (See *Soviet Physics, J.E.T.P.*, 1958, **6**, 620.)
9 Low, W., and Shaltiel, D., 1958, *J. Phys. Chem. Solids*, **6**, 315.
10 Ludwig, G. W., Woodbury, H. H., and Carlson, R. O., 1958, *Phys. Rev. Let.*, **1**, 295.
11 Manenkov, A. A., and Prokhorov, A. M., 1957, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **33**, 1116. (See *Soviet Physics, J.E.T.P.*, 1958, **6**, 860.)
12 Morigaki, K., Fujimoto, M., and Itoh, J., 1958, *J. Phys. Soc. Japan*, **13**, 1174.
13 Sanadze, T. I., 1957, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **33**, 1042. (See *Soviet Physics, J.E.T.P.*, 1958, **6**, 801.)
14 Tucker, R. F., 1958, *Phys. Rev.*, **112**, 725.
15 Weger, M., and Low, W., 1958, *Phys. Rev.*, **111**, 1527.
16 Zverev, G. M., and Prokhorov, A. M., 1958, *J. Expt. Theor. Phys.* (*U.S.S.R.*), **34**, 1023. (See *Soviet Physics, J.E.T.P.*, 1958, **7**, 707.)

EXPERIMENTAL INVESTIGATIONS OF THE IONOSPHERIC E-LAYER

By B. J. ROBINSON
Radiophysics Laboratory, C.S.I.R.O., Sydney, Australia†

CONTENTS

	PAGE
§ 1. Introduction	242
§ 2. Early investigations	242
§ 3. Routine penetration-frequency measurements	246
3.1. Identification of the penetration frequency	246
3.2. Diurnal variation of the penetration frequency	248
3.3. Variation of penetration frequency with solar activity	251
3.4. Seasonal variation of the penetration frequency	253
3.5. Variation of the penetration frequency with latitude	254
3.6. Influence of the magnetic current system	255
3.7. Other influences	257
§ 4. Measurements of the height of the E-layer	257
4.1. Routine measurements of group height	257
4.2. Special measurements	259
§ 5. Determination of the distribution of free electrons	260
§ 6. Direct measurements in the E-region using rockets	262
§ 7. The fine structure of the E-layer	264
§ 8. Behaviour of the E-layer during eclipses	268
§ 9. The effect of solar flares on the E-layer	269
§ 10. Formation of the E-layer	271
§ 11. Concluding remarks	275
Acknowledgments	276
References	276

Abstract. This paper discusses the recent information gained about the E-layer of the ionosphere by various experimental techniques. The basic information is provided by penetration-frequency measurements made from routine pulse soundings, giving a general picture of the variation of the ionization during the day, with season and with latitude, as well as the effect of changing solar activity and of the magnetic current system. Special radio-reflection experiments to measure height changes have yielded important results, particularly concerning tidal movements. The detailed structure of the layer has also been deduced from high-resolution pulse soundings. Independent information about the structure of the layer has been determined directly using rockets, and shows marked differences from the structure deduced from ground observations.

Observations with high-resolution ionosondes reveal that the reflections from the E-region are often very complex, and this introduces uncertainties in the identification of the characteristics of the normal E-layer. A brief discussion is given of observations of fine-structure in the layer.

Rocket measurements of solar radiations and of the atmospheric composition have made it possible to predict the distribution of ionization in the E-layer; this prediction is in quite good agreement with the distribution determined from accurate soundings.

Observations of changes in the ionization during solar flares and radio fade-outs are described.

---

† Present address: Sterrewacht, Leiden, Netherlands.

§1. Introduction

The major part of our knowledge about the earth's upper atmosphere has been deduced from radio measurements, although comparatively recently more direct measurements have been made possible by the use of rockets. Because of their overriding control of long-distance radio communication the strongly ionized F-layers above 150 km have been studied in very great detail. The less dense ionization below 150 km (taken for convenience as the upper boundary of the E-region, its lower limit being arbitrarily defined as 90 km) has attracted far less experimental and theoretical attention, and there is still some qualitative uncertainty about its behaviour.

The general features of the ionization have been determined from measurements of the maximum electron density in the layer, the bulk of these measurements having been made for communication purposes from variable frequency pulse soundings. Special radio-reflection experiments have given more detailed information about the structure of the layer, while measurements of the height of the layer have revealed tidal movements within the E-region. From accurate 'group height' measurements made over a range of frequencies it has been possible to calculate the detailed distribution of the ionization with height (up to the level of maximum ionization); such measurements are particularly valuable for a study of the processes by which the electrons are produced and disappear. Most important information about the distribution of ionization has also been gained by rocket measurements, which extend beyond the level of the maximum to heights where little can be determined by soundings from the earth. During rocket flights measurements have also been made of the atmospheric composition and of the intensity of ultra-violet and x-ray radiation from the sun; this has enabled considerable progress to be made in understanding how the layer is formed.

A vast amount of new experimental material has been collected during the International Geophysical Year 1957/1958, and it appears to be appropriate to take stock of the present position before the analysis of the new data begins. So particular stress has been laid on the areas where the information is incomplete, or where there is disagreement between various experimental results or their interpretation.

An account of the early radio investigations and of the general features of the ionosphere is to be found in the books of Mitra (1952) and Rawer (1953). A short review of the results of routine E-layer measurements has been published by Appleton (1955). More general information about the properties of the upper atmosphere may be found in the reviews by Bates (1949) and Weekes (1954).

§2. Early Investigations

The existence of an electrically conducting layer in the upper atmosphere had been suggested by Balfour Stewart's studies of the earth's magnetism in 1882, while in 1902 both Kennelly and Heaviside postulated that an ionized layer would account for the observed transmission of long radio waves over large distances. Direct experimental evidence for the existence of a conducting layer was provided a quarter of a century later by the frequency-change experiment of Appleton and Barnett (1925), the first pulse-soundings by Breit and Tuve (1925), and the angle of incidence experiments of Appleton and Barnett (1925) and of Smith-Rose and

Barfield (1926). It was soon found that waves above a certain frequency penetrated the Kennelly–Heaviside layer to be reflected from another, more densely ionized layer above. The first detailed studies of the lower layer—termed the 'E-layer' by Appleton—were those of Appleton and Naismith (1932); they estimated the frequency for which the waves penetrated the layer, which gave a measure of the maximum density of the charged particles causing the reflection (Appleton 1931). Comparison could then be made with the theory developed by Chapman (1931) for the ionization of an idealized atmosphere by monochromatic radiation from the sun. The observed penetration frequencies were markedly controlled by the sun, reaching a maximum near noon and falling to low values during the night. Observations during eclipses (Appleton 1930, Henderson 1933) made it certain that the principal ionizing agency was electromagnetic radiation from the sun.

Measurement of the polarization of the reflected waves (Appleton and Ratcliffe 1928) and the observation of appreciable magnetic double refraction at medium frequencies (Appleton and Builder 1933) showed that the reflections were mainly due to electrons. Appleton and Naismith (1935) measured the penetration frequencies for the two components of the reflected wave produced by magnetic double refraction; comparison with the magneto-ionic theory developed by Appleton (1927, 1932) and Hartree (1929) showed that the reflection conditions at medium latitudes corresponded to the 'quasi-transverse' approximation to the theory, which indicated that the effect of collisions between electrons and gas molecules was small at the level of maximum ionization.

E-layer measurements have always been closely compared with the predictions of Chapman's theory, for information could then be gained about the physical properties of the atmosphere above 100 km. When the electrons disappear by recombination the density N at the peak of a simple Chapman layer is described approximately by

$$\frac{dN}{dt} = q_0 \cos\chi - \alpha N^2 \qquad \ldots\ldots(1)$$

where q_0 is the maximum rate of electron production when the sun is at the zenith, χ is the solar zenith angle, and α is the recombination coefficient. When N attains its maximum value near noon we have

$$\alpha N^2 = q_0 \cos\chi. \qquad \ldots\ldots(2)$$

Appleton and Naismith (1935) measured the maximum electron density at noon as χ changed with season, and found that it obeyed equation (2). Kirby and Judson (1935) studied the diurnal variation, and from its marked symmetry about noon concluded that the rate of recombination was rapid, so that an equilibrium between the rates of recombination and loss of electrons existed well away from noon. During most of the day their measurements were in good agreement with equation (2). Appleton (1937) noted that the diurnal variation was very symmetrical about noon, and estimated that the recombination coefficient was greater than 10^{-8} cm^3 sec^{-1}, an unexpectedly large value. A careful study of the diurnal variation in winter by Best, Farmer and Ratcliffe (1938) revealed the same high degree of symmetry. An example of the diurnal variation found by

Best, Farmer and Ratcliffe is reproduced in figure 1; from these measurements the value of α was found to be at least $1{\cdot}2 \times 10^{-8}$ cm^3 sec^{-1}. From the symmetry of the variation observed at Huancayo and Washington, Hulburt (1939) found that α would need to be 2×10^{-8} cm^3 sec^{-1} or greater.

Appleton, Naismith and Ingram (1937) and Hulburt (1939) found that the seasonal variation of noon ionization agreed closely with equation (2) for stations at latitudes ranging from that of Huancayo (12°S) to Tromsö (70°N). However, Hulburt's work showed that the noon ionization in a given season decreased more rapidly with increasing latitude than predicted by (2).

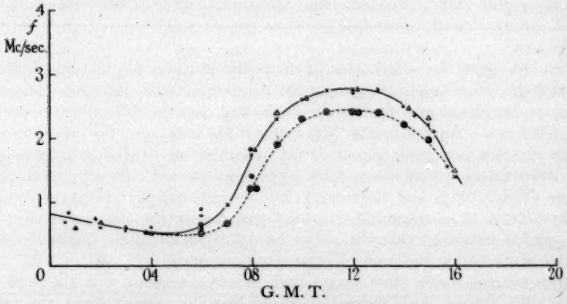


Figure 1. Diurnal variation of E-layer penetration frequencies, 15–18 December, 1936. (*From* Best, Farmer and Ratcliffe 1938).

As the sunspot cycle had progressed from 1932 to 1938 several workers had noted that the intensity of the ionization in all the ionospheric layers was following the rise in solar activity. Appleton and Naismith (1940) reported that at Slough the intensity of the ionization of the E-layer had shown both general and detailed correspondence with the sunspot number from 1931 to 1938, but that the relationship between the two appeared to differ for the rising and falling portions of the solar cycle.

Most of the early experimenters had noted that there was not always a single layer of ionization below the F-layers, although the exact nature of the extra layers was not determined with certainty. Schafer and Goodall (1933) observed a layer intermediate in height between the E- and F-layers, with "a well-defined maximum of ionic density". The group height of the echoes from this layer ranged from 130 to 180 km, in winter. Appleton (1933) verified that the intermediate layer had been observed in Britain, more frequently in the early morning. He suggested that it be termed the E^{II}-layer, following the F-region nomenclature. Martyn and Green (1933) and Ratcliffe and White (1934) observed reflections from the intermediate layer at sunrise and sunset in measurements of the group height on a fixed frequency. Such experiments gave little information about the behaviour of the layer during the rest of the day.

Within the next few years experimental techniques were greatly improved, and it became possible to make soundings over a continuous range of frequencies. This led to interpretations of the structure of the main layers which were less ambiguous than when the frequency was varied in steps. An interesting discussion of the continuous soundings was given by Gilliland (1935). Many of his records show several strata between the E- and F-layers; he pointed out that the multiplicity of penetration frequencies led to confusion in identifying the penetration of the E-layer, particularly when the structure of the echoes changed in the half-hour between his measurements. The intermediate layers were observed more frequently in winter than in summer.

Appleton and Naismith (1935) also observed an intermediate stratum which did not show group-retardation effects at its penetration frequency; they termed these reflections 'abnormal E'. The stratum was found to be very common in summer, and could show a small range of frequencies for which reflections were observed simultaneously from the stratum and the F-layer. The 'abnormal E' reflections were discussed further by Appleton, Naismith and Ingram (1937, 1939) and by Appleton and Naismith (1940). These workers also reported the occurrence of two or three strata between the E- and F-layers; in such cases the ionization associated with the 'abnormal E' reflections appeared to form in a sequence, beginning as a small ledge in the base of the F-layer whose group height fell steadily, its maximum electron density increasing and its thickness decreasing.

Harang (1937) reported that an intermediate layer was often observed in the morning at Tromsö, while a thinner, higher stratum was observed in midsummer. Best, Farmer and Ratcliffe (1938) found two distinct layers in winter in England, the two penetration frequencies varying in a regular manner following the altitude of the sun. The behaviour of the two penetration frequencies is shown in figure 1. In another investigation Best and Ratcliffe (1938) found as many as three distinct layers below the F-layer, and under undisturbed conditions the penetration frequencies of these three layers (separated by a few hundred kilocycles per second) varied regularly throughout the day. If the reflections were slightly disturbed these strata were difficult to observe.

In the early investigations very few attempts were made to measure the height changes in the E-layer, for accurate height measurements were difficult to make. By differentiating the echo pulse, Halliday (1936) achieved an accuracy of measurement to $\pm 0{\cdot}5$ km, though he could not resolve two pulses if their difference in group height was less than 5 km. Measurements in different seasons showed that the group height of E-layer echoes was lower in winter than in summer, a surprising result which has been contradicted by later work. Appleton and Weekes (1939) measured the group height on a fixed frequency (to the nearest 0·5 km) and detected a lunar semi-diurnal tidal variation in the height of the E-layer, of amplitude close to 1 km, which reached its maximum about three-quarters of an hour before lunar transit. At ground level the lunar tidal oscillation is minute, the amplitude at Greenwich being 0·0088 mm.

Our basic knowledge of the behaviour of the E-layer was well founded before normal scientific work was interrupted in 1939. During the war the number of routine ionospheric sounding stations grew rapidly in response to the needs of

military communication. This has led to an important change in the type of investigations which have since been carried out in connection with the E-layer. A large amount of effort has been devoted to the analysis of hourly measurements made from routine soundings by relatively unskilled personnel, the measurements being primarily intended for communication predictions. To an unfortunate extent this has led to isolation of research workers from the original records, so that some qualitative features of the reflections and instrumental limitations have not been fully appreciated. The number of well-designed experiments which have been executed form a very small part of the post-war investigations of the E-layer. We shall now review the results of the many analyses of routine penetration frequency measurements, and then proceed to consider the results of other experiments.

§3. Routine Penetration-Frequency Measurements

3.1. *Identification of the Penetration Frequency*

In all routine soundings of the ionosphere the delay time of a short radio-frequency pulse is recorded as a function of frequency. Close to the frequency where the waves can penetrate a thick layer the pulses show a very long delay

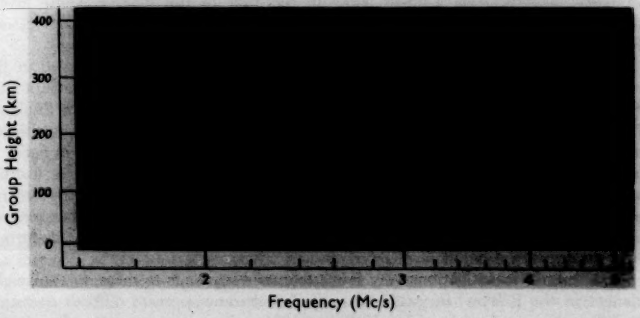


Figure 2. Vertical incidence pulse sounding: 16^h, April 20, 1958. E-layer penetrated at 3·3 Mc/s.

because they must traverse a considerable region where their group velocity is very small. If the layer were many wavelengths thick the delay at the penetration frequency would ideally become infinite. In practice the echo becomes weak before the delay time becomes very great, due mainly to the effects of deviative absorption and dispersion. Despite the weakening of the echo the retardation associated with the penetration of a layer is a very marked feature of an ionospheric sounding record, such as that reproduced in figure 2. However, the presence of deviative absorption means that the accuracy of penetration frequency measurements depends on the power of the sounder and on the local noise level; methods of estimating where penetration occurs on records such as that of figure 2 have been discussed very fully by Bibl, Busch, Rawer and Suchy (1955).

Not all records are of the simple type discussed above, where there is clearly a lower (E) layer and a higher (F) layer. Figure 3 shows a summer record where the penetration of the layer is incomplete. After some retardation has occurred the delay then decreases until the reflections cease and the echo from the F-layer appears. The intermediate stratum, for which the delay decreases with frequency, has been termed the 'abnormal E-layer' by Appleton and Naismith (1935). The penetration is rarely complete during the summer months, but the nature of the 'abnormal' reflections is not yet known. In routine measurements the frequency at which the cusp appears on the record (the maximum retardation) is measured as the penetration of the normal layer; if the retardation is very small an extrapolation of the trace is sometimes made (see Hanson, Hagg and Fowle

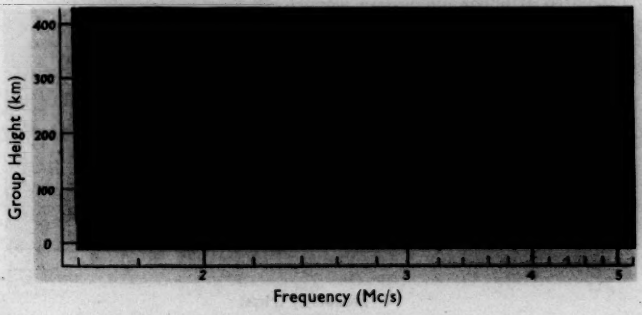


Figure 3. Vertical incidence pulse sounding: 08^h, April 18, 1958. Incomplete penetration of E-layer.

1953, Wright, Knecht and Davies 1957). The 'cusp' frequency has a diurnal variation similar to the frequency measured for complete penetrations; however, the details of the variation of the cusp frequency (Robinson 1957) show that it does not provide a very accurate measure of the maximum electron density in the layer.

Another common type of record is reproduced in figure 4, where several maxima are penetrated before the echo from the F-layer appears. Routine procedures (see the references cited above) do not provide an unambiguous means of deciding which penetration (if any) corresponds to that of figure 2. The most common procedure is to take the first penetration as that of the E-layer and to regard the extra penetrations as those of 'E_2' and higher layers. However, this classification is not adopted if the retardation at the first penetration is too small; Bibl (1951) takes the E-layer penetration to be the first with a retardation greater than 15 km, but this varies from station to station. Some workers have introduced further terminology, referring to E_0, E_1, E_{1b}, E_2 or F_0 layers (see, for example, Dieminger and Hoffmann-Heyden 1952, Rastogi 1954). The clarity of the penetrations depends markedly on the sensitivity and resolution of the sounder

used, and much of the confusion of nomenclature and lack of physical understanding of the multi-penetration records can be traced to instrumental limitations.

Records of reflections from the E-region are often more complicated, sometimes bearing no resemblance to any of the three main types discussed above. In some the trace may be obscured by spreading of the echoes; in others the height of the echo changes discontinuously as the frequency changes, as though a subsidiary layer had been penetrated without any group retardation. Examples of complex reflections have been reproduced by Wright, Knecht and Davies (1957). The caution shown in deciding whether a normal penetration can be identified in these cases varies from station to station and is again markedly influenced by the quality of the recording.

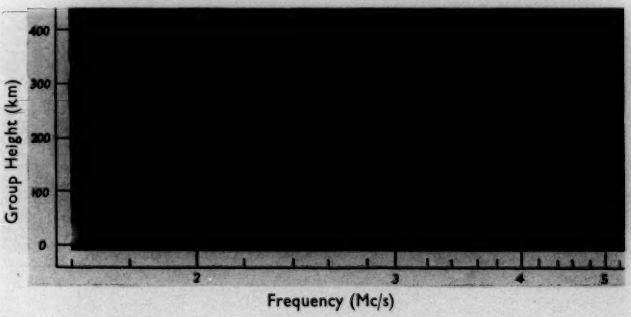


Figure 4. Vertical incidence pulse sounding: 14^h, December 18, 1957. Several maxima apparent in E-region.

The hourly measurements from routine stations show a large day-to-day scatter, and for most purposes only monthly median data are used. Bibl (1951) has found that a classification based on the amount of group retardation present leads to a marked reduction in the scatter of the daily measurements, as compared with that when the internationally recommended definition is used. Further evidence, to be discussed in § 8, makes it appear that much of the scatter in routine measurements does not represent real variations in the maximum electron density of the E-layer. Robinson (1957) has also shown that the errors which can occur in routine measurements are not always random.

3.2. *Diurnal Variation of the Penetration Frequency*

The hourly routine measurements provide a general indication of how the ionization varies during the day, and from this it is possible to deduce some information about the physical processes which occur in the E-region.

Most workers since the war have analysed the routine measurements in a way suggested by the equilibrium approximation to Chapman's theory, equation (2)

above. Departures from this expression have been treated by fitting the measurements with a law of the form

$$N_m = N_0 \cos^{2n} \chi. \qquad \ldots\ldots(3)$$

Values of the index n different from $\frac{1}{4}$ then show that there is some change from the behaviour of an ideal Chapman layer.

Waldmeier (1944) found that at Kochel n was close to $\frac{1}{4}$, notably different from the Chapman value suggested by the pre-war work. Tremellen and Cox (1947) determined for a large number of stations an average value of 0·31, with variations from 0·28 to 0·34. Using data for the years 1940–1948 from Huancayo, Watheroo, Freiburg and Slough, Harnismacher (1950) also found values of n from 0·29 to 0·36. For Canadian stations Scott (1952) noted that n was near 0·33 at moderate latitudes, but was accurately 0·25 to the north of the auroral zone; for stations within the auroral zone the daytime E-layer was found to be relatively insensitive to the solar elevation. With measurements collected from up to 50 stations Menzel and Wolbach (1955) reported values of n between 0·27 and 0·30. In table 1 are collected values of n (for the diurnal variation) for stations covering a wide range of latitudes.

Table 1. Dependence of E-Layer Ionization on Solar Elevation $N_m = N_0 \cos^{2n} \chi$

Station	Latitude	Diurnal variation		Seasonal variation[c]	
		n	$N_0 \times 10^{-4}$	n'	$N_0' \times 10^{-4}$
Resolute Bay	74·7 N	0·24[a]	16·6	—	—
Fairbanks	64·9 N	—	—	0·246	12·2
Baker Lake	64·3 N	0·24[a]	15·6	—	—
Churchill	58·8 N	0·13[a]	17·9	0·206	13·9
Fort Chimo	58·1 N	0·13[a]	—	—	—
Gt. Baddow	51·7 N	0·285[b]	12·5	0·270	12·6
Portage la Prairie	49·9 N	0·235[a]	18·1	—	—
Ottawa	45·5 N	0·27[a]	17·8	0·232	13·0
Washington	39·0 N	0·303[b]	14·5	0·284	14·1
San Francisco	37·4 N	0·305[b]	14·9	0·278	13·3
Baton Rouge	30·5 N	0·313[b]	15·6	0·244	12·6
Maui	20·8 N	0·322[b]	16·6	0·279	13·8
San Juan	18·4 N	0·324[b]	16·7	0·201	13·7
Trinidad	10·6 N	0·328[b]	17·2	0·332	14·5
Huancayo	12·0 S	0·329[b]	17·1	0·352	15·1
Brisbane	27·5 S	0·316[b]	15·9	0·278	13·1
Watheroo	30·3 S	0·314[b]	15·6	0·248	12·9
Canberra	35·3 S	0·308[b]	15·0	0·274	13·4

[a] Scott (1952): data for 1949–50.
[b] From expression by Harnismacher (1950): sunspot number $R = 0$.
[c] Allen (1948): $R = 0$.

For most of these stations the maximum electron density varies more rapidly throughout the day than is the case for a Chapman layer. Thus if the variation given by equation (2) is fitted to the noon measurement, the values at other times of day will lie below the theoretical curve. It should be noted that the earlier measurements reported by Appleton (1937) and Best, Farmer and Ratcliffe (1938)

fell above a $\cos^{3/2}\chi$ variation fitted to the noon values. This discrepancy remains unexplained; it is important to note that there has been no significant improvement in the accuracy of measurement since the pre-war experiments were performed.

A value of the index n different from $\frac{1}{2}$ does not call for any drastic modification of the assumptions of Chapman's theory. Since the layer is lowest at noon, any influence which increases the maximum density when it is low (or vice versa) could explain a larger value of n. An increase in either the scale height or recombination coefficient with height could produce this effect (Nicolet and Bossy 1949, Gledhill and Szendrei 1950, Appleton and Lyon 1955, Shimazaki 1956). However, observation of the penetration frequency does not yield any additional information about the scale height.

The penetration frequency measurements show a high degree of symmetry about noon (Harnismacher 1950, 1955), as revealed by the pre-war measurements. For a Chapman layer this can occur only if there is an equilibrium between the rates of production and loss of electrons throughout the day. Such an equilibrium requires a rapid rate of recombination; with a slower rate of recombination equation (1), § 2, shows that the ionization would display a time lag relative to the variation given by equation (2). If a time lag can be measured it is possible to estimate the value of the recombination coefficient, and Appleton (1937, 1953) has suggested two simple ways of determining the value of α from a time lag. However, any influence not allowed for in equation (1) can upset an estimate of α, and Appleton has made it clear that his methods give only an 'effective' value of the recombination coefficient, which he represents by α'.

Seaton (1948), Baral and Mitra (1950) and Appleton (1955) have investigated the departures from symmetry by comparing the ionization at times equally spaced from apparent noon. The values found for α' are not constant throughout the day; α' rises steadily as the times considered are spaced farther and farther from noon. The change in α' is far greater than any change in the actual recombination coefficient needed to account for the departure of the index n from its Chapman value. The approximate fit of the diurnal variation to equation (2) shows that the value of q_0/α can only vary by a small amount throughout the day. It is most improbable that q_0 (depending on the incoming photon flux and a reference value of the scale height) could vary by a considerable factor throughout the day. As Appleton (1955) has pointed out, it is even less likely that q_0 varies in such a manner as to keep q_0/α constant if α exhibits the large variations found in α'.

With a specially designed sounder Robinson (1957) made more accurate penetration frequency measurements which show that the values of α' given by the same method are fairly constant over the central portion of the day. At medium latitudes the values found for α' begin to rise rapidly four or five hours from noon; farther still from noon there is no 'recombination lag', the afternoon ionization being less than at a corresponding morning hour. This behaviour cannot be explained by any variation of the actual recombination coefficient with height. So although the E-layer behaves in a Chapman-like fashion over the middle part of the day some influence not included in the theory makes itself felt well away from noon, acting at medium latitudes to increase the maximum electron density in the early morning or to decrease it in the late afternoon. This influence

has not yet been identified, although there is no lack of likely physical influences: movements of ionization and photodissociation of the molecular oxygen in the E-region are both known to occur. The possibility of tidal effects will be discussed in § 3.6.

3.3. *Variation of Penetration Frequency with Solar Activity*

The pre-war investigations had shown clearly that the ionization of the E-layer is closely related to the degree of solar activity. This is the major long-term change in the E-layer, and must be allowed for before other changes can be discussed.

Various methods have been adopted to eliminate the terrestrial influences of time of day and season. Appleton and Naismith (1940) had used $f_0^4 \sec\chi$ as an index of the intensity of the ionizing radiation, the value for a Chapman layer; Allen (1946) extrapolated the penetration frequencies for a given hour and season to zero sunspot number, and then used the ratio of the observed penetration frequency to the corresponding extrapolated value as an index of changes in the ionization, raising this ratio to the fourth power to give a measure of the flux of radiation; Scott (1952) used the empirical relationship (3) discussed in the previous section to find the penetration frequency for an overhead sun, using values of n determined separately for each station. All these procedures implicitly assume that neither the scale height nor the recombination coefficient varies with the solar cycle or with season.

All workers have found that the corrected penetration frequencies are approximately linearly related to the relative sunspot number R, the agreement being better when long-term averages or running averages are considered. If we write f_0 for the value of the penetration frequency for overhead sun, f_x for its value, when $R = 0$, the relations found between f_0 and R can be written as follows:

Tremellen and Cox (1947): $f_0^{3\cdot2} \sec\chi = f_x^{3\cdot2}\,(1 + 0{\cdot}00735R)$;

Allen (1948): $f_0^4 \sec\chi = f_x^4\,(1 + 0{\cdot}0097R)$;

Harnismacher (1950): $f_0 \sec^n\chi = f_x + (0{\cdot}01\text{–}0{\cdot}007\cos\phi)\,R$,
which for latitude $\phi = 40°$ gives
$f_0^{3\cdot3} \sec\chi = f_x^{3\cdot3}\,(1 + 0{\cdot}0092R)$;

Scott (1952): $f_0 \sec^n\chi = f_x(1 + 0{\cdot}0018R)$,
which with $n = 0{\cdot}28$ (at Ottawa) gives
$f_0^{3\cdot6} \sec\chi = f_x^{3\cdot6}\,(1 + 0{\cdot}0065R)$.

None of this work has reflected the pre-war observation (Appleton and Naismith 1940) that the dependence of f_0 on R was different in years of rising solar activity from that in years of falling activity. Naismith (1943), for Slough measurements from 1931 to 1942, gave the following expressions:

(i) for rising R: $f_0^4 \sec\chi = f_x^4\,(1 + 0{\cdot}033R)$;

(ii) for falling R: $f_0^4 \sec\chi = f_x^4\,(1 + 0{\cdot}027R)$;

which in either case shows a much greater sensitivity to variations in solar activity than reported by more recent workers.

The values found for the sub-solar penetration frequency at medium latitudes for zero sunspots are :

Naismith (1943) :	$f_z = 2{\cdot}8$ Mc/s ;
Tremellen and Cox (1947) :	$f_z = 3{\cdot}3$ Mc/s ;
Allen (1948) :	$f_z = 3{\cdot}3$ Mc/s ;
Harnismacher (1950) :	$f_z = 3{\cdot}4$ Mc/s ;
Scott (1952) :	$f_z = 3{\cdot}2$ Mc/s.

There is much better agreement between these values, except for the pre-war measurements.

It is important to note that the close relationship between f_0 and R appears only for fairly long-term averages. Short-term variations are often not related, and it is not clear whether this is because of random scatter in the daily measurements, unidentified terrestrial variations, the possibility that the sunspot number serves only as an index of the general level of solar activity, with no detailed relationship to changes in the intensity of the ionizing radiation, or whether the radiation depends on the phase of the solar cycle as well as on the instantaneous value of R. Allen (1946) noticed that during a period of low solar activity the monthly median f_0 was depressed for some months when there was great magnetic activity. Beynon and Brown (1951) examined the daily noon measurements at Slough for 1947–1950, and found that there was a 27-day periodicity in the penetration frequency which was almost in phase with the 27-day periodicity in sunspot number during the same periods. However, the sensitivity of f_0 to these short-term changes in R was only half that shown for long-term averages.

Allen (1948) has made a detailed comparison of the ionization of the E-layer with a variety of solar phenomena : sunspots, bright hydrogen and calcium flocculi, faculae, and the intensity of the green coronal line (5303 Å). The monthly mean values of all these factors show a fairly detailed agreement, but the variations of each one have a time delay of a fraction of a month with respect to the sunspot variation. The E-layer ionization lags the sunspot variation by about 0·4 of a month, intermediate between the delay for calcium flocculi (0·31 month) and faculae (0·56 month). This has led Allen to conclude that the radiation producing the E-layer comes from the solar chromosphere and is associated with the appearance of flocculi and faculae. In an earlier paper Allen (1946) has also noted a degree of correlation between the daily values of E-layer ionization and daily floccular character.

Denisse and Kundu (1957) have compared the E-layer measurements at Fribourg between 1947 and 1953 with the intensity of 10 cm solar noise (as measured in Ottawa), with the area of sunspots and with the intensity of the coronal line emission. They find that the monthly mean ionization is most closely correlated with the intensity of the 10 cm solar noise ; an inspection of their diagrams reveals that there is still not a correspondence in detail.

There are some observations which indicate that the ionization may not bear a one-to-one correspondence with the instantaneous value of any other solar index.

Minnis (1957a, 1958b) found that the localized sources and general background of the ionizing radiation invoked to explain E-layer behaviour during eight eclipses had varied differently with the solar cycle. The rate of change of the total radiation was twice as great as that for the background radiation; the background radiation was found to have varied by a factor of 2 : 1 during the present solar cycle. Beynon and Brown (1951, 1957) have also shown that the relation between changes in the E-layer and in sunspot number is different for short-term changes and long-term averages. They suggest that this indicates a change in the background radiation during the solar cycle.

In the case of 10 cm solar flux measurements, the use of high resolution aerials has made it possible to separate changes in the background radiation from those due to relatively short-lived active areas. Minnis (1957a) has pointed out that the ratio of the change in background to total radiation for the 10 cm solar flux is similar to that detected in the E-layer by eclipse observations. When combined with the work of Denisse and Kundu, referred to above, it appears that the 10 cm flux may be a useful single index of changes in the ionizing radiation. Minnis (1958b) has also found that the distribution of sources of ionizing radiation on the solar disc deduced from observation of an eclipse (at sunspot minimum) was strikingly similar to the distribution of 10 cm noise sources observed at Ottawa with a high-resolution aerial.

Information gained from rocket measurements (to be discussed in §10) and observations of sub-structure in the E-layer (§7) suggest that the E-layer is formed mainly by x-rays from the solar corona, with some contribution by ultra-violet radiation from the chromosphere. It is unlikely that these radiations behave similarly throughout the solar cycle. The lack of detailed correlation of the total ionizing radiation with the 10 cm solar flux (also from the corona) is then not surprising.

3.4. *Seasonal Variation of the Penetration Frequency*

When a correction has been made for changing solar activity, it is of interest to see whether the penetration frequency at one station depends only on the solar elevation as this changes through the seasons.

Tremellen and Cox (1947) detected a small annual variation of the penetration frequency for a fixed solar elevation, the values being increased in November, December and January and decreased in May, June and July for stations in both hemispheres. Harnischmacher (1950) found the same world-wide effect in the values of N_0 deduced from equation (3) for different months, the annual variation being of the order of $\pm 5\%$. He pointed out that the variation of the sun–earth distance from perihelion in January to aphelion in July would cause a change in N_0 of the same order of magnitude. Menzel and Wolbach (1955) also noted that N_0 was a maximum all over the earth in November, December, January and February, though they were of the opinion that this was due to the influence of sporadically occurring strata and other E-layer perturbations "prevalent during the winter months". (About two-thirds of the stations they considered were in the northern hemisphere.) Looking again at penetration frequencies for a fixed solar elevation, Appleton (1955) noted at Slough a maximum in December, while Beynon (1955) reported that for Slough the morning values showed a subsidiary

maximum in June, not present when afternoon values were considered. Scott (1952) found a conflicting result for Canadian stations, where N_0 was a minimum in December and a maximum either in June or, in some years, at the equinoxes.

Tremellen and Cox (1947) also claimed that there was a small seasonal effect causing a reduction of the penetration frequency in local summer. Neither Harnismacher nor Menzel and Wolbach noted any seasonal effect of this kind, while Beynon's subsidiary maximum at Slough is just the opposite effect.

Some workers who have determined the index n in equation (3) from the diurnal variation have found that annual influences are also reflected in the values of n deduced at different times of the year. Menzel and Wolbach, and Beynon, noted that n was a maximum (about 0·3) at the equinoxes, with minima (about 0·27) in summer and winter. Harnismacher found only a very small annual variation in n, following the variation in N_0. Scott's values of n and N_0 rose and fell together month by month, so that his values of n also vary in a way not found at other stations.

Allen (1948) has determined the variation of N_m at different stations with the seasonal changes of $\cos\chi$ for a given hour of the day. Having first removed the effect of the changing sun–earth distance, he found that the measurements were fitted closely by an expression similar to equation (3); the values of the index varied from 0·20 to 0·35 at different stations, the higher values occurring nearer the equator. The mean index for fourteen stations was 0·268. Allen noted that the values determined from the noon measurements tended to give slightly higher values for the index than those determined from the measurements at 09^h and 15^h; the investigation of this difference by Beynon and Brown (1956) is discussed in § 3.6.

In table 1 we have added Allen's seasonal index (which we shall denote here by n') to the values of the diurnal index n for the same stations. We find that for all stations away from the equator n is greater than n'. Thus while N_m is quite accurately a function of $\cos\chi$ for both variations, it is more sensitive to the rapid changes through the day than to the slow changes during the year. Since n is greater than n' in both hemispheres, we have here an effect depending on the local season. Relative to the value of N_m at a given hour in local summer, the value at the same hour in winter is found to be greater (by about 6%) than when the sun is at the same elevation on a summer day. Since the value of n (diurnal) varies very little between summer and winter, this means that N_0 is greater in local winter than in summer. This effect was not reported by Harnismacher, or by Menzel and Wolbach; it may be the 'small seasonal effect' referred to by Tremellen and Cox, but as they did not provide any numerical estimate a comparison is not possible.

At equatorial stations n' appears to be greater than n, so that N_m is more sensitive to the slow changes in $\cos\chi$ throughout the year than the rapid changes during the day. The annual range of $\cos\chi$ being small, it would be valuable to have an estimate of the accuracy of n' to see whether the amount by which it exceeds n is significant.

3.5. *Variation of the Penetration Frequency with Latitude*

For a given season and local time the solar elevation is different at different latitudes and, if the atmosphere were the same above all points of the earth, the

measurements at different stations would again be in accord with equation (3). Tremellen and Cox (1947) found that for a fixed solar elevation N_m decreased as the latitude increased. The values of N_0 which Allen (1948) deduced from the seasonal variation also decreased with increasing latitude in both hemispheres, the values near 60°N being 20% below those found at the equator. In addition Allen's values of n' tended to be greater at lower latitudes. Comparing the diurnal variation at different stations, Harnismacher (1950) found that both N_0 and n decreased away from the equator, and suggested expressions for these for latitudes less than 60°. Menzel and Wolbach (1955), Appleton (1955) and Shimazaki (1957) all confirm that there is always more ionization towards the equator. However, Scott (1952) examined the data from the same stations considered by the previous workers (but for different years) and found very little variation of N_0 with latitude. This result has been supported by the work of Sato (1955). It is very difficult to see how such conflicting results can be reached by workers analysing data from similar routine observations. The extensive measurements made during the International Geophysical Year 1957–8 should help to resolve the question of the variation with latitude.

3.6. *Influence of the Magnetic Current System*

Three groups of workers have recently reported that the quiet-day magnetic current system exerts an appreciable influence on the ionization in the E-layer.

Appleton, Lyon and Pritchard (1955) determined the delay of the maximum E-layer ionization relative to noon (for the month of June in 1949 and 1953) for eleven stations covering a wide range in latitude. The delay varied with latitude, and Appleton *et al.* suggested that the pattern of the variation of delay time with latitude is related to the quiet-day current system ; this can influence the ionization by introducing vertical drift effects when it flows across the earth's magnetic field. At a latitude of 40°N their results give a delay of more than 20 minutes after noon, while at the equator the ionization reaches its maximum a few minutes before noon. Later Appleton, Lyon and Turnbull (1955) showed that there were abnormalities in the daily variation of N_m and of the minimum observed equivalent height h_{min}' of the E-layer at Slough in summer. They pointed to a marked depression of N_m about 1000–1100 hours local time, which had the effect of displacing the time of maximum ionization to about 40 minutes after noon ; the minimum equivalent height was also shown to be about 2 km lower in the middle of the morning than at the corresponding time after noon. The fall in h_{min}' is consistent with the direction of flow of the overhead current system before noon (Martyn 1953), while the depression of N_m requires the downward drift velocity to decrease with altitude at this latitude. Appleton *et al.* point out that if such drift effects are reversed at low latitudes it is possible to account for an increase of N_m (for fixed χ) towards the equator. The accurate measurements of the 'phase height' (see §4) of the E-layer by Findlay (1951a) also showed that the minimum heights occurred 10 to 20 minutes before apparent noon.

It should be noted that Harnismacher (1950, 1955) has also determined the times for maximum ionization in the E-layer. Using monthly median values, he found that the axis of symmetry for the diurnal variation corresponded very closely (± 2 min) with the apparent noon at the station. This is stated to be the

case for Huancayo, Watheroo, Fribourg, Slough, Dakar, Nha Trang, Djibouti and Lwiro, and the variations of the time of maximum compared with the local mean noon clearly followed the equation of time. Harnismacher found a fixed time delay at some stations corresponding to the difference between the nominal time of recording and the time at which the exploring frequency approached the E-layer penetration. A plot of the 'equation of time' deduced from eight years' data for Huancayo has been published by Rawer (1955). The centre of symmetry of this curve is about 7 or 8 minutes before the nominal mean noon; as the published Huancayo value for any hour is actually the value for a time nearer 10 minutes past that hour, Rawer's curve shows that the ionization of the E-layer at Huancayo (12°S) reaches its peak value 2 or 3 minutes after apparent noon. (Huancayo lies almost exactly on the 75°W meridian, so no further time correction is necessary.)

By making frequent soundings throughout summer days, when the penetration of the E-layer is not complete, Robinson (1957) has shown that the critical or 'cusp' frequency (figure 3) is not a reliable measure of the maximum ionization density (particularly if only hourly soundings are made). The summer delay-time measurements should therefore be regarded with caution. In other seasons more reliable measurements of the maximum density can be made, but the overhead current system is weaker than in summer. In spring at medium latitudes there is still a depression of the group heights by 1 to 2 km over a major part of the morning which is presumably due to the solar tide (lunar tidal effects having been removed). However, careful measurements of the diurnal variation of penetration frequency (Robinson 1957) show that it only suffers a depression of 30 kc/s in the mid-morning (a change of 2% to 3% in the maximum electron density). This did not produce any measurable delay of the maximum ionization (± 5 minutes).

Beynon and Brown (1956) have produced evidence of a different sort to suggest that the magnetic current system influences the ionization in the E-layer. They have analysed published monthly median values by the same method as used by Allen (1948)—taking the seasonal variation at a given hour to find the seasonal index n' (see § 3.4). The value found for n' is sensitive to any annual or seasonal influences, and the effect of the ellipticity of the earth's orbit is most marked in Beynon and Brown's results. Using noon data for stations lying within a restricted range of longitude for the years 1952–1954 they found that n' showed a most striking dependence on latitude, with singularly low values at Watheroo and Tokyo, near the latitudes of the foci of the current system. Interpreting these results as due to the influence of vertical drift requires that: (i) an overhead current towards the west be associated with a decrease in the penetration frequency, (ii) the singularly low values of n' be found at latitudes such that the current focus is located on opposite sides of the station in summer and winter.

In support of this interpretation Beynon and Brown show that the values of n' based on the data at 1500 hours, when the E–W magnetic current is small at all latitudes, are larger than for the noon data, have no singularities, and vary in latitude in the way found by Allen. For one of the stations (Tokyo) giving a singularly low n' for noon data, the values of n' for each hour of the day were computed. The value found varied with the hour chosen, being lowest for the

1100 hour data ; this is also close to the time when the E–W component of the overhead current reaches its maximum value.

Shimazaki (1957) has examined published data in yet another way. For a fixed local time at an equinox he determined the variation of N_m with latitude, all the available stations being considered together. Curves of the form

$$N_m = N_0 \cos^{2n''} \chi \qquad \ldots\ldots(4)$$

are fitted to the results at each hour. The values found for N_0 and n'' are different for the different hours considered, as can be seen from table 2. The curves described by equation (4) fit the measured points more closely for the afternoon hours than for the morning hours, the morning points for low latitude stations tending to lie above the empirical curves. Shimazaki suggests that this is due to vertical drift effects ; he finds that at all latitudes these are small for the afternoon hours, but for low latitudes become considerable in the morning. Shimazaki (1958) also finds that at the solstices the ionization is no longer symmetrical about the latitude where $\chi = 0$ but has its maximum value nearer the equator, as found by Appleton (1955). He attributes this to the greater strength of the magnetic current system in the summer hemisphere.

Table 2. Dependence of E-Layer Ionization on Latitude $N_m = N_0 \cos^{2n''} \chi$

(The table gives values of n'')

Month	09^h	10^h	11^h	12^h	13^h	14^h	15^h	Mean value
March	0·332	0·335	0·325	0·312	0·332	0·322	0·315	0·325
June	—	0·252	0·248	0·260	0·260	0·282	—	0·260
Sept.	0·338	0·338	0·335	0·322	0·338	0·340	0·340	0·335
Dec.	—	0·282	0·268	0·272	0·282	0·288	—	0·278

Data from Shimazaki (1958).

3.7. *Other Influences*

(i) Harnismacher (1955) has stated that when the effects of season, solar activity and the equation of time are removed there remains a small lunar influence. More details have been given by Rawer (1955), who has published a diagram showing that N_0 is increased at the time of the full moon. The published statements are confusing as regards the magnitude of this effect ; for some stations the change listed for N_0 is as large as 20%.

(ii) Allen (1946), as mentioned before, found that the monthly median values of N_m were depressed during some periods of marked magnetic disturbance near sunspot minimum. A similar effect for 5-day means had been noted by Appleton, Naismith and Ingram (1937).

§4. Measurements of the Height of the E-Layer

4.1. *Routine Measurements of Group Height*

In the pulse sounding method of Breit and Tuve the primary quantity measured is the total time of flight of the pulse to the ionosphere and back. This is usually converted to a group, or equivalent, height, which is half the distance the pulse would have travelled in the same time at the free-space velocity of light. The group height is always greater than the actual height at which the exploring pulse

is reflected, for within the ionosphere the group velocity is reduced below its free-space value.

Compared with the many investigations of the behaviour of the penetration frequency of the E-layer, the variations in the group heights of the layer have been almost neglected. This is mainly due to the difficulty of measuring group heights accurately, and the smallness of the regular variations. For example, routine measurements to the nearest 10 km at Watheroo were discontinued before the war because little diurnal variation of the group height could be found.

Most routine observatories measure the group height at the lowest frequency where echoes are recorded, a frequency which varies throughout the day and with season depending on the absorption suffered by the exploring pulse and on the level of interference. With a standard sounder it is not usually possible to measure this height (denoted by $h_{\min}'$) to better than the nearest 5 km, and many stations measure to the nearest 10 km. Rawer and Argence (1953) have published a diagram showing the spread of the measured heights throughout the day for one month. An attempt was made to measure $h_{\min}'$ to the nearest 2 km, but the scatter is such as to mask all but the slightest trace of a diurnal variation. Sato (1955) has published a study of the routine values of $h_{\min}'$ which demonstrates that rounding off to the nearest 10 km hides any variation of height with latitude.

With a high resolution sounder it is found that a single, clean echo is not often returned from the E-region. On many occasions the echoes have a complex structure which either seriously reduces the accuracy of height measurement or introduces uncertainty in the identification of the echo from the normal E-layer.

On about a third of all occasions the echoes are spread over a considerable range of heights, and the errors of measurement on a high-resolution sounder can be three or four times as great as when there is a single echo. Spreading of the echo is the chief limitation to accurate height measurements.

Often a sharply reflecting stratum appears a little below the heights at which the E-layer is usually formed. Because of the variability of its occurrence and the rapid changes which occur in its upper frequency limit, this stratum is commonly referred to as sporadic E (E_s). McNichol and Gipps (1951) have suggested that it be called 'constant height' E_s in order to avoid confusion with other uses of the term sporadic E. Sometimes the difference in height between the echoes from E_s and from E is small, and unless the sounder has good height resolution the jump may not be noticed. At frequencies less than 1·6 Mc/s reflections similar to these occur very commonly, the group heights being somewhat lower—95 to 100 km. These reflections were often recorded by Halliday (1936), and may be related to the E_s reflections seen on higher frequencies. In the height range 95 to 100 km a patchy, partially reflecting stratum is also observed. Naismith (1954) has suggested that it is due to reflections from a 'meteoric' E-layer, though Gregory (1956) has since questioned the association of these reflections with meteoric ionization.

At times a stratum can appear to form within the E-layer, partially obscuring the normal echoes, with a sudden jump in group height at its upper frequency limit. These echoes are sometimes introduced by a redistribution of the ionization in a normal layer (Robinson 1957), although at times their upper frequency limit can rise rapidly far above the normal penetration frequency.

4.2. *Special Measurements*

Many measurements have been made of the group height on a fixed frequency. Such measurements (to the nearest 0·5 km) were used by Appleton and Weekes (1939) to detect a lunar tidal variation in the height of the E-layer, of amplitude close to 1 km. Weekes (1948) made an approximate analysis of the diurnal variation of group height to determine the scale height in the E-region, finding values ranging between 6·5 km and 10 km. Findlay (1951 a) measured the change in group height with season on a fixed frequency. From September to March the group height varied in a manner consistent with Chapman's theory, with a scale height of 8·4 km; but the summer measurements were lower than predicted by the theory. Findlay also measured the changes in the phase of the wave train reflected from the E-layer, which enabled him to detect changes in the phase path

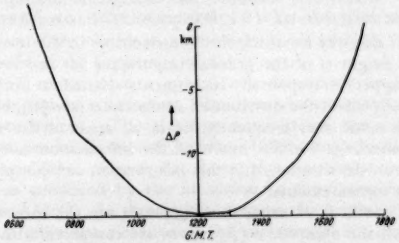


Figure 5. Changes in phase path ΔP during August 10, 1949, after removal of the lunar tide from the observations. (*From* Findlay 1951a.)

of the waves of less than a wavelength, i.e. less than a few hundred metres in his experiment. On the few days when the diurnal variation of phase path could be followed continuously there was an asymmetry about noon, which was greatly reduced if the effects of the lunar tide were removed. One of Findlay's (corrected) curves is reproduced in figure 5. The diurnal variation agreed well with that for a Chapman layer with an atmospheric scale height of about 6 km, a much smaller value than was believed to be possible. There remained a slight asymmetry about noon, which showed itself through the fact that the morning fall of the layer was more rapid than the afternoon rise, so that the phase path was a minimum about 10 or 20 minutes before noon. This asymmetry is of the same type as that found in h_{min}' at Slough by Appleton, Lyon and Turnbull (1955).

Martyn (1948) has used published values of h_{min}' (measured to the nearest 10 km) to study tidal effects in the E-region. Matsushita (1953) has similarly analysed published values of the group height of E_s echoes (measured to the nearest 10 km). Thomas and Svenson (1955) made fixed-frequency measurements which were capable of an accuracy to $\pm 0{\cdot}2$ km. They point out that during the day it is very difficult to identify the reflections from the base of the E-layer because of the presence of sporadic E; at Brisbane it was only possible to identify the E-layer reflections with certainty on 15% of the records. Thomas and Svenson therefore measured the height of E_s echoes during the night to

detect the lunar tide. All these workers have found a lunar semi-diurnal tide of amplitude 0·5 to 0·7 km, but find different values for its phase. Gazzard (1958) has made a tidal analysis of the echoes from the 'abnormal' E-layer. He found a tide of amplitude 1·5 km, different again in phase; he interprets these differences as due to a change in the tidal oscillation with height.

Accurate height measurements would provide important information about the solar tide which is thought to modify the maximum electron density in the E-layer. The experimental work required would be difficult, but the effects reported in §3.6 require a tide of amplitude comparable with the lunar oscillation.

A great amount of information has been obtained from the study of long waves reflected from the lower E-region; an excellent summary of this work has been published by Waynick (1957).

§5. Determination of the Distribution of Free Electrons

Knowledge of the way in which the free electrons in the lower ionosphere are distributed with height is of the greatest importance for the investigation of the physics of the upper atmosphere. Instruments carried in rockets can provide information directly about the distribution of electrons (see §6), but such measurements cannot be made very frequently, or in all areas of the world. However accurate $h'(f)$ soundings contain much of the information necessary for determining the electron distribution, and this information can complement the rather infrequent rocket measurements.

The simplest methods of using the $h'(f)$ data are the comparison, or model, methods, in which the observed $h'(f)$ curves are compared with those computed for assumed distributions of simple shape. These methods have been discussed by Becker (1956) and by King (1957). Appleton (1937) compared E-layer $h'(f)$ curves with those for a parabolic model, neglecting the influence of the earth's magnetic field, and found that the scale height in the region was approximately 10 km. Similar comparisons by Weekes (1948) gave values of the scale height ranging between 9 and 23 km, with a mean value of 17 km. Whale (1950) compared automatically recorded $h'(f)$ curves with those for model distributions, without field corrections. The $h'(f)$ curves were rarely as simple in shape as those for the models, and the comparison procedure became very complicated. Though Whale attempted to ignore the peculiarities in the shapes of the $h'(f)$ curves, he found that the range of parameters required for the best-fitting models was large, and so the values of scale height were very uncertain.

Waves reflected by the E-layer are close enough to the gyro-frequency for the effect of the earth's magnetic field on the group height to be very important. Wartime calculations by Cox (unpublished) showed that the neglect of magnetic field effects led to an overestimate of the scale height by a factor of about 2 when model methods were used. Shinn and Whale (1952) have also shown that the thickness of a parabolic layer is overestimated by about 80% when the method of analysis neglects the magnetic field.

When the layer shape is complex, or when several layers are present, it is more satisfactory to determine the electron distribution directly by inverting the integral equation for the group height. Some of the analytic and numerical methods for determining the electron distribution have been reviewed by Becker

(1956), while King (1957) has discussed all the methods at present available, and the many factors which affect the accuracy attainable.

Grace (1951 a) has used a numerical method due to Kelso (1952), which ignores the magnetic field, to deduce the electron distribution from $h'(f)$ curves measured manually with a special sounder. The published results for one day showed that the height of the peak of the layer was almost constant throughout the day, completely different from the variation for a Chapman layer. Measurements of the recombination coefficient at heights below the peak, by a direct extension of the method of Appleton (1937), revealed that the coefficient decreased rapidly with height.

Whale (1951 a) also attempted to determine the electron distribution. He used a fully automatic sounder, but the accuracy with which he could measure the group height was seriously limited by changes in the width of the recorded echo when its amplitude changed as the frequency was varied. (This effect has been discussed by Lyon and Moorat (1956).) Whale tried to develop a method to include the effect of the earth's magnetic field in his analysis, but did not make any systematic study of the properties of the E-layer.

Robinson (1957) constructed a special sounder which made it possible to measure group heights to within one kilometre when the reflections were free from natural disturbances. The electron distributions were determined by a numerical method similar to that of Jackson (1956) which includes the effect of the earth's magnetic field. It was also found that an allowance had to be made for the retardation produced by the ionization in the D-region, particularly at frequencies below 2 Mc/s. In these measurements the height of the peak ionization was found to follow the elevation of the sun throughout the day in approximate agreement with the behaviour of a Chapman layer.

It is difficult to explain the difference between the results of Grace (1951 a) and Robinson (1957). It is not likely to result from the allowance for D-region retardation in the latter analysis (important only at frequencies well below the penetration frequency), nor could it result from the fact that Robinson included the effect of the magnetic field in his calculations and Grace did not. The only feasible explanation seems to be that in Grace's manual measurements, at 100 kc/s intervals, it would have been more difficult to recognize disturbances of the $h'(f)$ curve. The height of the peak cannot be determined very reliably, for the errors of measurement are greatest when the echo is greatly retarded and suffers deviative absorption.

Robinson has compared the shape of the measured electron distributions with Chapman's theory, and finds that to a first approximation the diurnal variation of the distribution is in agreement with the theory. A detailed comparison showed that the scale height increased with altitude, the values determined being fairly consistent with those determined from rocket measurements and observation of meteor trails. These measurements are compared in table 3. Below 110 km the E-layer measurements give appreciably lower scale heights than the other methods. Wark and Stone (1955) measured the Doppler width of the 5577 Å line in the radiation from the night sky and found that it corresponded to a temperature of 231°K; later rocket measurements by Berg, Koomen, Meredith and Scolnik (1956), and by Heppner and Meredith (1958) established that this line was emitted from a thin layer between 94 and 100 km. These results correspond to a scale height

of 7·1 km near 100 km in a mixed, undissociated atmosphere (mean molecular weight 29); dissociation of the molecular oxygen would increase the scale height deduced from the temperature, up to a limit of 8·5 km for total dissociation.

Table 3. Scale Heights in the E-Region

(1)	(2)	(3)	(4)
100	—	6·3	7·0
102·5	5·25	6·5	7·3
105	5·4	6·8	7·5
107·5	6·2	7·1	—
110	7·4	7·4	—
112·5	8·6	7·7	—
115	9·4	8·25	—
117·5	10·0	8·9	—
120	—	9·4	—

(1) Altitude (km); (2) E-layer observations (km), from Robinson (1957): all values ±0·5 km; (3) rocket measurements (km), from Horowitz and LaGow (1957); (4) meteor observations (km), from Kaiser and Evans (1955).

The increase of the scale height with altitude will modify the diurnal variation of the peak density in the layer (see §3.2). The level of the peak does not fall below 108 km (for overhead sun), but the scale height gradient above this level would make the diurnal index n in equation (3) about 0·30.

Robinson (1957) determined the recombination coefficient from the variation of the layer shape with time. The values of the coefficient showed an apparent decrease with height; the rate of decrease is much smaller than that reported by Grace (1951a), but is greater than could be produced by possible changes of temperature with height (an upper limit to the change in temperature is provided by the change in scale height). The lag produced by recombination is very small, and the accuracy reached so far in determining the electron distributions is barely adequate for a measurement of the recombination coefficient. The mean value found by Robinson was $(1{\cdot}5 \pm 0{\cdot}5) \times 10^{-8}$ cm^3 sec^{-1}. The behaviour found by Grace and by Robinson could be produced by the movement of ionization, by changes in the dissociation of molecular oxygen, or by the existence of more than one ionizing process. No comparable change is found in the values of q_0/α given by accurate penetration frequency measurements (relating to heights above 108 km).

§6. Direct Measurements in the E-Region using Rockets

Since the war, captured V2 rocket missiles and developments from them have many times reached heights above the E-layer, and have been able to carry instruments to make direct measurements of the properties of the atmosphere and of parts of the solar spectrum which cannot be observed from the ground. The problems associated with these experiments have been enormous, but it is now possible to make some measurements with considerable accuracy. For the moment we shall restrict discussion to the determination of the electron density in the E-region.

So far three different methods of measuring the electron density using rockets have been employed, each susceptible to errors of interpretation and each yielding

different results. The most accurate method has been devised by Seddon (1954 b): by comparing the Doppler shift for one c.w. signal which is influenced by the ionospheric electrons with a reference signal of much higher frequency, it is possible to determine the refractive index for the lower frequency as a function

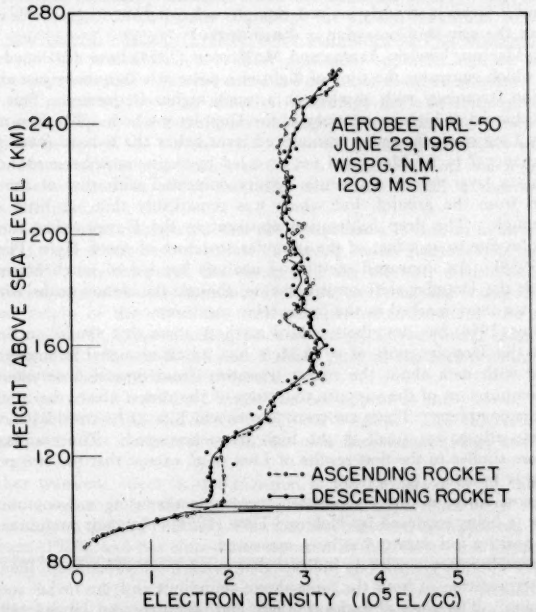


Figure 6. Ascent and descent measurements of electron densities with Aerobee-Hi rocket. (*From* Jackson and Seddon 1958.)

of height. The electron density can then be deduced from the dispersion equation. When the two magneto-ionic modes are separated by suitable receiving aerials, comparison of the refractive indices for the two modes provides a most valuable check on the measurements. Rocket measurements by Seddon (1954 a), Jackson (1954), Seddon, Pickar and Jackson (1954), Jackson, Kane and Seddon (1956) and Jackson and Seddon (1958) have revealed that the ionosphere above 100 km is a strongly ionized continuum, showing no very distinct layers. The marked features on the $h'(f)$ curves which identify the E- and F_1-'layers' are associated with quite minor peaks on an essentially monotonic increase of ionization. An example of rocket measurements is reproduced in figure 6.

Sounding the E-layer from the ground yields information only up to the peak of the layer, and provides no direct measurements of the distribution of the ionization up to the level where the electron density next exceeds that at the peak. However, from observation of the retardation of echoes from the F-region, Hollingworth (1935), Bibl (1950) and others had inferred that the region between the E- and F-layers is strongly ionized, though such experiments give little information about the way this ionization is distributed.

Lien, Marcou, Ulwick, Aarons and McMorrow (1954) have developed another method which compares the time of flight of a pulse at a frequency just above the penetration frequency with another on a much higher frequency. This method has not achieved as high an accuracy as the Doppler method. The first measurements by Lien *et al.* showed a pronounced layer below the E-layer (with penetration frequency 2 to 2·5 Mc/s, and not revealed by their ionospheric sounder) and above this a layer whose maximum density coincided with that of the E-layer measured from the ground, but which was remarkably thin—as little as 3 km right through. The deep bifurcation apparent in the E-region has since been found to be due to an effect of the irregular structure of the E-layer (Pfister and Ulwick 1958). An improved method of analysis has led to much better agreement with the Doppler-shift measurements, though the minor peaks within the E-region are more marked in the pulse-delay measurements.

Berning (1954) has described a third method, somewhat similar to Seddon's, in which the Doppler shift of a 74 Mc/s c.w. tracking signal is measured and compared with data about the rocket trajectory from optical observations, and from a computation of the vacuum trajectory of the rocket above the level where burn-out is complete. There are many errors which must be considered, and the ionospheric effects are small at the high frequency used. The results of two flights were similar to the first results of Lien *et al.* except that the two peaks are considerably broader.

The possibility of direct ionization measurements using a Langmuir-probe technique is being explored by Hok and Dow (1954). In their preliminary tests only the positive ion density has been measured.

These preliminary results do indicate that much more effort has to be devoted to the interpretation of both the ionospheric soundings and the rocket results for the E-region. The work of Jackson (1956) and of Pfister and Ulwick (1958) has shown that the usual standard of $h'(f)$ sounding falls far below the standard required for studying the E-layer, and will need to be improved considerably to match the results of rocket measurements.

§7. The Fine Structure of the E-Layer

There can often be marked stratification of the E-layer, and the pre-war observations of this have already been discussed (§2). In recent years some high-resolution ionosondes have been constructed, and a few workers have again stressed the great complexity that can be shown by reflections from the E-layer. Bibl (1951) has sketched about a dozen types of reflections commonly observed at Fribourg. Up to the present date the complex features of the reflections have not been given any detailed study, for the experimental difficulties involved are formidable. A small degree of spreading of the echoes is sufficient to blur out

many small details, while the appearance of sporadic reflecting strata in or below the layer can hide them completely.

(*a*) *Transient effects.*

Some of the subsidiary maxima which are apparent on $h'(f)$ curves have a relatively short life. With hourly soundings the second maximum may appear on only one record. Whale (1951 b) made soundings at four-minute intervals and observed that sometimes a subsidiary maximum appeared high on the $h'(f)$ curve, and moved steadily down the curve in a period of about 30 minutes. Minnis (1956 b) also observed the same effect on soundings made at frequent intervals during a solar eclipse. In these cases the higher penetration frequency appears to be related to that of the normal layer before the transient subsidiary maximum appears. However, the appearance of such a disturbance on the $h'(f)$ curve should make us cautious about assuming that there is still a 'normal' layer present.

Some workers have suggested that the disturbance is due to a wave motion in the layer, while others have postulated a moving source (localized) of ionization. Robinson (1957) has computed the changes in the electron distribution during the movement of a subsidiary maximum. In the case analysed it was found that the actual height of the subsidiary maximum remained constant, the apparent movement on the $h'(f)$ curve taking place as the layer rose (under solar control) through this level. It was found that a shear wind in the region could account satisfactorily for the disturbances observed.

At high latitudes a subsidiary maximum can appear to move down the $h'(f)$ curve, and then to reverse and move up the curve again. This behaviour can be accounted for by the horizontal movement of a disturbance (Hagg and Hanson 1954).

Another transient effect in the E-region is the formation of an 'abnormal' E-layer in a sequence which also begins with the appearance of a subsidiary maximum high on the $h'(f)$ curve. This was first observed by Appleton, Naismith and Ingram (1939), and has since been reported by McNichol and Gipps (1951), Whale (1951 b), Bibl (1951), Saha and Ray (1955) and Rastogi (1956). McNichol and Gipps have suggested that this phenomenon be termed 'sequential E_s'. The group heights of the reflections above the first maximum are observed to decrease steadily; the amount of group retardation observed at the second penetration decreases at the same time until there is only a discontinuous jump in group height (similar to that shown in figure 3). At low latitudes the sequence continues further, resulting in the formation of low-level 'constant height' E_s, and may also produce severe spreading of the E-layer reflections.

Because of the fall in group height during the sequence many workers have spoken of "ionization moving down from the F-region to the E-region". Robinson (1957) computed the changes in the electron distribution during such sequences and found that there was no downward movement of ionization, the sequence being produced by a drastic redistribution of the ionization in the E-layer. Though there was a large fall in group height the ionization was found to be transported upwards (over a range of about 10 km).

(*b*) *Behaviour of the 'abnormal' E-layer.*

The work discussed in the previous section has made it clear that the form of an $h'(f)$ curve can give a most misleading impression of phenomena occurring in the E-region. Because the reflections produced in figure 3 show no group retardation at their upper frequency limit they have been termed 'abnormal' (Appleton and Naismith 1935), or 'sporadic', in contrast to 'normal' layers which show group retardation at their penetration frequency (Bibl 1951, 1953). However, the type of sequence described above shows that there is no sharp line of demarcation between the two cases. An examination of the electron distributions (Robinson 1957) has also shown that small changes in the distribution could produce large changes in the form of the $h'(f)$ curve, and that there often was no striking physical difference associated with the presence or absence of group retardation. The shape of the $h'(f)$ curve depends on the distribution in the layer below as well as on that in the reflecting layer.

The upper frequency limit of the 'abnormal' E reflections can behave in a most erratic fashion, and this has lent support to the terms abnormal or sporadic. However, on many occasions the upper frequency limit of these reflections behaves in a regular fashion—as appears to be the case in the study by Best, Farmer and Ratcliffe (1938) of their summer 'threshold' frequencies. Becker and Dieminger (1950a, 1951) have also noted the regular variation of the limiting frequency on many occasions and have queried the use of the term 'abnormal', suggesting that it should be classified as a normal, 'E_2', layer. They find that the limiting frequency is usually 200 to 300 kc/s above the cusp frequency, with a day-to-day scatter of 10%; the minimum group height varies greatly from day to day, but this is not surprising in view of the small penetration frequency difference between the layers.

(*c*) *Persistent stratifications.*

In winter in England, Best, Farmer and Ratcliffe (1938) found that when their records were free from disturbances two penetration frequencies could be observed which persisted throughout the day. Both frequencies behaved in a regular fashion with changes in the solar elevation (see figure 1). On very quiet winter days Best and Ratcliffe (1938) found that three distinct penetrations could be followed for many hours.

Rastogi (1954) has found that at Ahmedabad there can be three layers below the F_1-layer which behave in a regular fashion. The uppermost layer is the 'abnormal' E-layer and Rastogi, like Becker and Dieminger (1951), has classified it as the E_2-layer. The two lower layers show group retardation at their penetration frequencies and so are of the same type as those observed by Best, Farmer and Ratcliffe. Rastogi reports that this multiple structure is most clearly seen shortly after sunrise, and is least clear during the noon hours; in the afternoon it is not as clear as in the morning due to the effects of interference and absorption. The formation of these layers at sunrise has been described by Rastogi (1956). For the months of February, March and April 1953, the multiple structure was observed on about 80% of the records. For these months each of the three penetration frequencies exhibited a diurnal variation of the form $f_0 \propto \cos^n \chi$, where

the index n had mean values of 0·375, 0·37 and 0·385 for the first, second and third penetrations.

Robinson (1957) has studied the multiple structure above Cambridge, England. In winter clear records showed two penetrations with group retardation throughout the day. The cusp on the $h'(f)$ curve associated with the lower penetration was very marked in the early morning; it was smallest during the noon hours and increased again towards sunset, but did not become as marked as it had been in the morning. In the morning and late afternoon a third, lower penetration, with only a small amount of group retardation, could be followed.

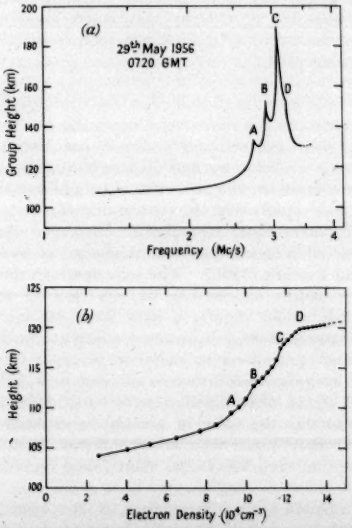


Figure 7. (*a*) E-layer sounding showing multiple structure at Cambridge, England. (*b*) Electron distribution computed from the measurements shown in (*a*).

The degree of stratification varied considerably from day to day. At the equinoxes two, and sometimes three, penetrations were observed on clear records in the early morning; as the sun's elevation increased the cusps on the $h'(f)$ curve decreased steadily until, towards noon, they could no longer be seen; in the late afternoon the lower penetrations appeared again. In summer the stratifications were observed only near sunrise, the cusps corresponding to the lower penetrations disappearing within a few hours. Comparison of $h'(f)$ curves taken in different seasons at times giving the same solar elevation showed that the stratifications were similar in all seasons. At Cambridge these persistent stratifications

had usually become indistinguishable on the $h'(f)$ curve when the sun's elevation exceeded 25°.

The nature of the stratifications in the E-region cannot be determined from the meagre experimental results available. It should be remembered that even quite considerable retardation can be produced by a small ledge on the underside of a layer. In figure 7 is shown an $h'(f)$ curve measured in Cambridge, and the corresponding $N(h)$ curve; it is seen that small inflections on the $N(h)$ curve produce greatly magnified effects on the $h'(f)$ curve.

These phenomena merit much careful experimental work. A higher instrumental standard than is reached by routine sounders is necessary, and separation of the many effects which can occur requires soundings at frequent intervals. Computations discussed in §10 indicate that better experimental information about the structure of the layer might aid considerably in our understanding of the ionization processes concerned.

§8. Behaviour of the E-Layer during Eclipses

Eclipse measurements are the subject of a recent book—*Solar Eclipses and the Ionosphere* (Ratcliffe 1956)—and will only be briefly considered here. The observations provide the only available method of determining the distribution of the sources of ionizing radiation on the solar disc. In addition the changes during an eclipse occur far more rapidly than the normal diurnal changes, and recombination effects should be more easily measurable. However, the interpretation of eclipse measurements is by no means unambiguous, as has been stressed by Piddington (1951) and Ratcliffe (1956). The time at which the ionization reaches its minimum value is usually observed to be very close to totality; if the sun emitted uniformly this would require a very large recombination coefficient. However, there are features of the eclipse curve which are most easily interpreted by assuming that some of the ionizing radiation proceeds from localized bright areas, and when an irregular distribution is allowed it is difficult to obtain an independent estimate of the recombination coefficient. If all the ionizing radiation were cut off at totality the value of α could be estimated from the fall in ionization during the eclipse, but if there is any radiation from outside the visible disc there will be large uncertainties in the value found for α (see Minnis 1956a, Ratcliffe 1956).

Minnis (1956a) examined the observations of the same eclipse from two stations, and showed that to explain the results in terms of a non-uniform sun the value of α above both stations had to be given a value close to $1{\cdot}2 \times 10^{-8}$ cm^3 sec^{-1}. Bates and McDowell (1957) have queried the assumption of an irregular distribution of sources, and have suggested that more than one recombination process may occur. They have computed (for the case of two processes) eclipse curves which are similar to the observed ones. Minnis (1958a) has replied that real eclipse curves have features which can be taken to show the covering and uncovering of bright spots, and that it is not possible to explain the curve computed by Bates and McDowell in terms of a non-uniform sun. Although eclipse observations are difficult to perform and to interpret they are a most valuable addition to other methods of studying the E-layer, and are capable of providing information not available from other measurements.

Eclipse measurements have made it seem very probable that there are intense, localized sources of ionizing radiation on the solar disc, but the routine penetration frequency measurements do not show marked signs of their appearance and disappearance. In particular, the sources are found with disturbing regularity near the limb of the sun, and yet there are no marked signs of changes in the ionization as the sun rotates. Beynon and Brown (1951) examined the daily noon measurements at Slough for 1947–1950, and found only a very small 27-day periodicity. Minnis (1956a, 1957b) computed the mean value of $f_0^n \sec\chi$ (see § 3.2) from the routine hourly measurements for each day. The day-to-day variation in this quantity was considerably less than that of the value for any one hour, being smaller again than when the penetration frequency is determined by Bibl's definition. Minnis found that the way in which the daily average of $f_0^n \sec\chi$ varied in the 'control' periods about two eclipses provided evidence in favour of localized sources of ionizing radiation whose existence had been deduced from the measurements on the eclipse day. It would be very desirable to have more observations of this kind of day-to-day variation, with carefully identified penetration frequencies; it is also important that such measurements should extend over longer periods than covered by Minnis' work, preferably for several solar rotations. Unfortunately, such a series of measurements cannot be carried out at all times of the year; in summer when the penetration of the layer is incomplete no measurements of sufficient accuracy are possible; in winter the reflections from the normal layer are frequently obscured by low-level E_s. It is during the spring that the E-layer is freest from disturbances, so that a long series of reliable measurements may be obtained.

§ 9. The Effect of Solar Flares on the E-Layer

One of the terrestrial effects associated with the appearance, in H_α light, of a bright eruption in the solar chromosphere is the strong absorption of short-wave signals, propagated via the sunlit ionosphere, for a period of 30 minutes or more. These short-wave 'fade-outs' were first described by Mogel (1931) and Dellinger (1935, 1937), and many investigations have since shown that the chief effect is the production of excess ionization in the absorbing D-region. Because the waves reflected from the E-layer are greatly attenuated and their intensity usually falls below that due to local noise, it is very difficult to detect whether any changes have also occurred within the E-region. It is only during the return to normal conditions after the flare that the E-layer can again be observed.

Berkner and Wells (1937) noted that in the case of a class 3 flare the penetration frequency of the E-layer when the echo returned was about 200 kc/s (5%) greater than before the fade-out. Further observations of this type were reported by Berkner (1939). For most of the intervening period few observations of the changes in the penetration frequency of the E-layer have been made, principally because of the practice of recording once each hour, so that at the most only one observation is possible during the usual duration of a fade-out; many fade-outs would have run their whole course between soundings. Higgs (1942) reported an occasion when the penetration frequency of the E-layer during the recovery from a fade-out was increased by 15%. With recently developed high-power sounders the time for which the echoes are below the noise level is much shorter

than with most routine sounders. With a high-power sounder, Becker and Dieminger (1950b) observed a 30% increase in the penetration frequencies of both the E- and 'E_2'-layers (see §7(*b*)). During the fade-out associated with a class 3 flare, Dieminger and Geisweid (1950) observed a definite increase in the penetration frequency by 30%. Bibl (1951) also observed a definite case when the penetration frequency increased by more than 10% after a class 3 flare; on this occasion the penetration frequency remained above the mean for nearby days for more than three hours after the beginning of the fade-out. After making a careful examination of his records for May, June and July 1950, Bibl reported that every increase of the penetration frequency by more than 200 kc/s above the mean was coincident with a fade-out or a solar flare. Minnis (1957b) has also reported an increase in the ionization by 10% at the time of a small fade-out. No accompanying flare was observed.

Taubenheim (1957) has examined the records taken with the high-power sounder at Lindau from 1951 to 1956 (30-minute intervals) and has found that the majority of fade-outs were accompanied by an increase of the penetration frequency above the mean for six neighbouring days. He does not discuss the magnitudes of the increases observed, but seven specimen results show changes between 10% and 30%. In some of Taubenheim's examples it can also be noted that there is a very slow return to the mean condition after the fade-out, which he shows must be due to the persistence of the enhanced ionizing radiation after the visual effect has subsided.

Very few observations of the changes in the height of the E-layer have been published. The early work of Berkner and Wells (1937) showed that the minimum virtual height h_{min}' of the E-layer was 20 km higher immediately after the fade-out than before; however, if one plots the values of h_{min}' against the lowest frequency at which echoes were detected they reproduce the shape of the $h'(f)$ curve recorded when recovery was complete. There was thus no major change in the group heights. Other measurements of h_{min}' have been reported by Berkner (1939), but as the $h'(f)$ records were not published, no conclusions can be drawn about actual changes in the group heights.

Measurements of the changes in phase-height for waves of frequency near 2 Mc/s were made by Findlay (1947, 1951b). At the start of a fade-out he observed a rapid decrease in the phase height by about 0·5 km before the echoes became too weak to be recorded. When the echoes returned the phase-height was increasing slowly and rose by about 0·5 km to its stationary value. Either a decrease in the level of reflection, or an increase in the electron density below the reflection point, would produce a decrease in the phase path, but would cause opposite changes in the group path. Findlay was not able to measure changes in the group height directly because of changes in the width of the recorded trace when the echo amplitude decreased. So he estimated the change in group height from the changes in phase-height on two nearby frequencies. He reports that although there was always an increase in the group height, it was often not as great as the decrease of phase-height. Findlay nevertheless interpreted the near equality in magnitude of the changes in group and phase-heights as evidence that most of the excess ionization was produced below the reflection point for 2 Mc/s waves.

By using a medium-power sounder and recording in such a way as to minimize the errors due to changes in echo amplitude, Robinson (1957) was able to measure directly the changes in group heights during fade-outs. Waves reflected from the base of the normal E-layer were observed to decrease in group height by from 0·5 to 1 km ; on some occasions the waves were reflected from a sporadic stratum below the E-layer, and then no change could be detected to the measuring accuracy of $\pm 0{\cdot}5$ km. For the normal layer the drop in group height for frequencies well below the penetration frequency implies that ionization was produced in the base of the layer, in addition to the D-region ionization causing the absorption. The changes in the electron distribution were computed for two small fade-outs, and showed that the excess electron production was a maximum near 105 km, falling off rapidly above and below this level. Further observations of this type would be very useful, particularly if combined with measurements of the changes in phase-height. It would be interesting to know whether the large increases of penetration frequency usually associated with class 3 flares represent enhanced ionization near the level of the peak only or whether the increase is greater still at lower levels in the layer.

Chubb, Friedman, Kreplin and Kupperian (1957) made rocket observations of the solar radiation about 10 minutes after a small flare. Measurements were made of Lyman α radiation, which is thought to be responsible for the normal D-region ionization, but no increase above the level of L_α on neighbouring days was detected. However, a considerable increase in the intensity of x-rays in the 3 Å range was detected, suggesting that strong coronal excitation remained well after the visual flare.

§10. Formation of the E-Layer

The main physical interest of studies of the E-layer lies in the processes of electron formation and loss which occur, and in the information these yield about the properties of the earth's atmosphere in the vicinity of 100 to 130 km, and about the ionizing radiations emitted by the sun. Only a limited amount of information can be deduced directly from the E-layer observations (see the reviews of Bates 1949, and Weekes 1954) and progress has only been achieved by combining the primary observations with theoretical deductions and with other geophysical and solar measurements. The information gained in recent years by high-altitude rockets is most valuable, providing detailed information about the ultra-violet and x-ray spectrum of the sun, and about the properties of the atmosphere in and above the E-region. It is now possible to compute the contributions of various ionizing processes to the formation of the E-layer, and to compare such computations with the observed properties of the layer.

The information gained by rocket measurements up till 1953 has been discussed by Weekes (1954), and so we shall consider mainly the more recent measurements.

Spectrographic studies of the airglow had long revealed that molecular oxygen was partially dissociated above 100 km, and this is easily accounted for by the action of solar ultra-violet radiation between 1300 and 1750 Å. Friedman, Lichtman and Byram (1951), Havens, Friedman and Hulburt (1955) and Byram, Chubb and Friedman (1954) have made rocket measurements of the intensity of

the radiation near 1450 Å, and shown that it increases rapidly between 90 and 140 km. The measurements make it possible to determine the distribution of O_2. Byram, Chubb and Friedman (1954) have found that the distribution of O_2 between 108 and 130 km is very close to the diffusive distribution; at 110 km the measured density of O_2 is near 10^{11} cm^{-3}, about 75% of the oxygen being dissociated. The measured distribution calls for some comment, for observations of meteor trails and radio echoes have revealed wind systems likely to produce mixing of the gases in the E-region. Also Townsend, Meadows and Pressley (1954) and Meadows and Townsend (1958) have used a mass spectrometer mounted on a rocket to show that the ratio of the concentrations of argon and nitrogen is constant from 96 to 137 km. Mange (1955) has used rocket measurements of the neutral particle density to calculate that the rate of diffusive separation is very small at 110 km. The distribution of O_2 depends, however, on the relative rates of dissociation and recombination, as well as on diffusion and mixing. The expected distribution of O_2 has been discussed by many workers, and most exhaustively by Nicolet and Mange (1954) and Mange (1954, 1955); diffusion appears to be a faster process than dissociation, and much faster again than recombination, so that the final effect of mixing is to bring the distribution closer to that of diffusive separation than to that of dissociative equilibrium.

The other major atmospheric constituents, O and N_2, should be well mixed, and their distribution can be determined from the measurements of the total particle density by Byram, Chubb and Friedman (1956) and Horowitz and LaGow (1957).

Several radiations likely to produce ionization penetrate as far as the E-region. Early rocket measurements showed that there was a sufficient flux of soft x-rays above the E-region to produce all the ionization below. The ionization which could be produced by x-rays has been calculated by Bates and Hoyle (1948), Rawer and Argence (1953), Elwert (1953) and others. The absorption cross section of air in the 50–100 Å range is about 10^{-18} cm^2 (see Compton and Allison 1954), so if the scale height is about 8 km, the total particle density at the level of maximum absorption would be $1{\cdot}4 \times 10^{12}$ cm^{-3}. The measurements of Horowitz and La Gow (1957) show that this density is to be found near 110 km. Byram, Chubb and Friedman (1956) found that a total energy of 0·1 erg cm^{-2} sec^{-1} in the wavelength range 40 to 100 Å was absorbed between 110 and 130 km. This corresponds to an available photon flux of about 4×10^8 photons cm^{-2} sec^{-1}, which could produce most of the ionization in the E-layer.

Watanabe, Marmo and Pressman (1955) re-measured the first ionization potential of O_2 as $12{\cdot}07 \pm 0{\cdot}1$ ev, equivalent to the energy of a photon of wavelength 1027 Å. They pointed out that O_2 could then be ionized by the L_β line of the solar spectrum ($\lambda = 1025{\cdot}7$ Å). Watanabe *et al.* found that the total absorption cross section of O_2 for L_β was $1{\cdot}54 \times 10^{-18}$ cm^2, while that for N_2 was only 4×10^{-21} cm^2. With $H = 8$ km, the maximum absorption of L_β by O_2 would occur near the level where there are 8×10^{11} oxygen molecules cm^{-3}. The rocket measurements (Friedman 1955) show that this concentration would be found near 105 km. The energy available in L_β has been measured by Tousey (referred to by Waynick 1957) as 0·01 erg cm^{-2} sec^{-1}, equivalent to a photon flux of about 4×10^8 cm^{-2} sec^{-1}. The ionization yield of L_β in O_2 was measured by Watanabe

et al. as 58%, so that this process should make a considerable contribution to the ionization in the lower part of the E-layer.

Watanabe *et al.* also found that the absorption cross section of O_2 over the whole range from 850 to 1027 Å was greater than 10^{-18} cm², with strong bands where the cross section exceeded 5×10^{-17} cm². For N_2 the cross sections in this range were all less than 10^{-20} cm². Thus all radiation between 850 and 1027 Å will be absorbed between 105 and 120 km. The total flux in the range 1100 to 1300 Å has been measured by Byram, Chubb, Friedman and Gailar (1953), and by Johnson, Purcell and Tousey (1954). The energy flux in the continuum is much less than that in the line L_α (1215 Å), and so is not likely to produce appreciable ionization. O_2 has a low absorption cross section for L_α. Nicolet (1954) has shown that the energy flux in the Lyman continuum ($\lambda = 990$ Å) may be considerable, and this may also produce ionization in the E-region.

Recently Houston (1957, 1958) has reported detailed calculations of the production functions for these radiations, making an estimate of the energy contained in the Lyman continuum. Houston's values for the total particle density are derived ultimately from Havens, Koll and LaGow (1952); this latter measurement has since been modified (Byram, Chubb and Friedman 1956, Horowitz and LaGow 1957) so that Houston's densities for N_2 and O may be too high by a factor of 3. He has not published his computed production functions, but has calculated the electron distributions; this required a knowledge of the recombination coefficients for the three ions produced. For N_2^+ Houston adopts the laboratory value $\alpha = 10^{-8}$ cm³ sec^{-1} (Biondi and Brown 1949), so that the ionization of nitrogen contributes very little to the resultant electron density. The efficiency of change transfer to O_2 (Bates and Massey 1947) makes it probable that O^+ and O_2^+ have the same overall recombination coefficient. Houston has assumed that they are the same, and adopted the values of α at different heights proposed by Mitra and Jones (1954) from a synthesis of many radio measurements. The discussion in §§ 3, 5 and 8 has shown that estimates of α based on radio data are very uncertain; also the rate of decrease of α with height adopted by Mitra and Jones appears to be much too rapid.

Houston's calculations show a very marked bifurcation of the E-layer. Near sunrise two nearly separated layers are formed, the upper by the ionization due to soft x-rays, the lower due to L_β. The contribution from the Lyman continuum is comparable with that from L_β, but it has a rather broad maximum above the level of the maximum produced by x-rays. As the sun rises the separate layers begin to overlap (due to the smaller scale height of O_2), but even with the sun overhead there is a marked bifurcation.

Regular bifurcation of the E-layer under undisturbed conditions has been observed by Best, Farmer and Ratcliffe (1938), Best and Ratcliffe (1938) and Rastogi (1954). Robinson (1957) has observed that the stratification of the layer is more marked at low solar elevations, and that the separate layers merge rapidly when the solar elevation increases above 25°. He has also computed the electron distributions corresponding to the bifurcated $h'(f)$ curves; the degree of separation of the layers is found to be much less than that computed by Houston, and at the latitude of Cambridge was never observed to be as great as that computed for overhead sun.

The probable overestimate of the total particle density by Houston means that the layer formed by x-rays may be too high by about a scale height. So a calculation has been made (from the data given by Houston (1957)) to see the effect of dropping the production due to x-rays by about 8 km. The resulting electron distributions (in equilibrium) are shown in figure 8, for solar zenith angles of 0°, 65° and 80°. When $\chi = 0°$ the ionization due to L_β alone has its maximum near 104 km, but this is not resolved from the other contributions.

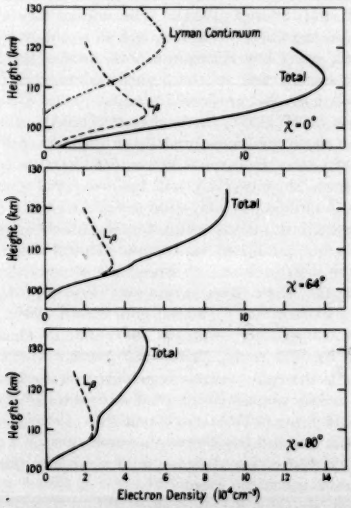


Figure 8. Computed ionization in the E-region due to x-rays, L_β and the Lyman continuum (data *from* Houston 1957). Electron distributions are shown for three values of the solar zenith angle χ.

When χ is greater than 60° the ionization due to L_β produces a small bump on the underside of the layer.

Electron distributions measured in Cambridge (Robinson 1957) when the solar zenith angle was close to the values used in figure 8, are shown to the same scale in figure 9. There is quite good agreement between the shapes of the observed and computed curves. The height of the layer agrees quite well with the calculations (the measured distributions do not extend right to the layer maximum because the penetration of the layer was not complete on the summer $h'(f)$ curves used). However, the model layer is somewhat thicker than shown by the measurements. Also the bifurcation of the model is rather low in the layer, compared with the observed subsidiary maxima. In figure 9 are also shown rocket measurements by Seddon, Pickar and Jackson (1954).

The uncertainties in the data used in Houston's computations could easily account for the lack of detailed agreement with the observations at Cambridge. The absolute values of the ionization density are in as good agreement as one could expect, for the ionization varies by more than 50% during the solar cycle, the rocket measurements of the flux of radiation could be out by 50%, and the values adopted for α are certainly too great, probably by a factor of 2. Further progress in theoretical and laboratory studies of the recombination processes would be most valuable.

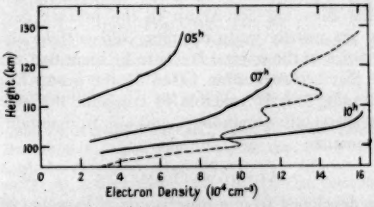


Figure 9. E-layer electron distributions at Cambridge (June, 1956) at times when the solar zenith angle was near the values shown in figure 8. The dashed curve shows rocket measurements (10^h, May 1954) at New Mexico by Seddon, Pickar and Jackson (1954).

Best and Ratcliffe (1938) and Robinson (1957) observed that there is sometimes a third stratification in the E-layer for large values of χ. Rocket measurements reveal no radiation or atmospheric constituent likely to produce this. Bates and Hoyle (1948) have suggested one process not considered by Houston, ionization by secondary photons emitted after an x-ray photon has removed an inner electron from an atom. Neither theoretical calculations nor the experimental observations are yet adequate for any comparison between them to be made.

§11. Concluding Remarks

Radio measurements alone have shown that the E-layer on its underside does not depart radically from an idealized Chapman layer. However, modifications are required in most of the assumptions of the theory. The scale height in the E-region certainly increases with height. Soundings from the ground and rocket measurements have given values of the scale heights which are in fair agreement above 110 km, but much below this level the soundings give smaller scale heights than revealed during rocket flights. The recombination coefficient appears to decrease with height, though the accuracy of measurement is not yet adequate to separate this completely from other small non-Chapman effects. Radio observations and rocket measurements also indicate that the assumption of a single ionizing process is not adequate to describe the layer. The main sources of ionization suggested by rocket measurements have been shown to account quite well for the observed ionization. Observations during eclipses have also shown that the ionizing radiation is not emitted uniformly from the solar disc.

Evidence has been obtained from routine penetration frequency measurements to suggest that the world-wide behaviour of the E-layer is influenced by the

over-head current system responsible for the quiet-day magnetic variations. The many measurements made during the International Geophysical Year 1957/58 should give valuable information about the world-wide variations. The interpretation of these variations would be greatly strengthened if accurate measurements of the changes in the height of the layer were made.

Many small effects have been revealed by the analysis of routine penetration frequency data, and by special experiments, which merit further attention and call particularly for further experimental work. However, development of our knowledge of the phenomena in the E-region will require a deeper level of qualitative understanding than has served up to the present day. The reflections from the E-region are usually quite complex, and current classifications do not allow the characteristics of the normal layers to be identified with certainty. It is important to note that a classification based on the group heights of reflections can be most misleading; it is valuable to compute the electron distributions corresponding to representative soundings, and also to maintain close comparisons with rocket measurements.

Acknowledgments

This report has developed from a programme of E-region investigations being conducted at the Cavendish Laboratory, Cambridge. I am very grateful to the Royal Society for the award of a Rutherford Memorial Scholarship to enable me to work at Cambridge.

It is a pleasure to acknowledge the assistance I have had in the preparation of this report from Mr. J. A. Ratcliffe and Dr. K. Weekes, of the Cavendish Laboratory. The records reproduced in figures 2, 3 and 4 were kindly provided by the Ionospheric Station, Netherlands P.T.T., Leidschendam. Figures 1, 5 and 6 have been extracted from papers by other workers, and acknowledgment for their use has been made in the text.

References

Allen, C. W., 1946, *Terr. Mag.*, **51**, 1; 1948, *Ibid.*, **53**, 433.

Appleton, E. V., 1927, *U.R.S.I. Rep.*, Washington; 1930, *J. Instn Elect. Engrs*, **66**, 872; 1931, *Nature, Lond.*, **127**, 189; 1932, *J. Instn Elect. Engrs*, **71**, 642; 1933, *Nature, Lond.*, **131**, 872; 1937, *Proc. Roy. Soc.* A, **162**, 451; 1953, *J. Atmos. Terr. Phys.*, **3**, 282; 1955, *Proc. Mixed Comm. on Ionosph.*, 4th Meeting, p. 14.

Appleton, E. V., and Barnett, M. A. F., 1925, *Proc. Roy. Soc.* A, **109**, 621.

Appleton, E. V., and Builder, G., 1933, *Proc. Phys. Soc.*, **45**, 208.

Appleton, E. V., and Lyon, A. J., 1955, *Physics of the Ionosphere* (London: Physical Society), p. 20.

Appleton, E. V., Lyon, A. J., and Pritchard, A. G., 1955, *J. Atmos. Terr. Phys.*, **7**, 292.

Appleton, E. V., Lyon, A. J., and Turnbull, A. G., 1955, *Nature, Lond.*, **176**, 897.

Appleton, E. V., and Naismith, R., 1932, *Proc. Roy. Soc.* A, **137**, 36; 1935, *Ibid.*, **150**, 685; 1940, *Proc. Phys. Soc.*, **52**, 402.

Appleton, E. V., Naismith, R., and Ingram, L. J., 1937, *Phil. Trans. Roy. Soc.* A, **236**, 191; 1939, *Proc. Phys. Soc.*, **51**, 81.

Appleton, E. V., and Ratcliffe, J. A., 1928, *Proc. Roy. Soc.* A, **117**, 576.

Appleton, E. V., and Weekes, K., 1939, *Proc. Roy. Soc.* A, **171**, 171.

Baral, S. S., and Mitra, A. P., 1950, *J. Atmos. Terr. Phys.*, **1**, 95.

Bates, D. R., 1949, *Mon. Not. R. Astr. Soc.*, **109**, 215.

Bates, D. R., and Hoyle, F., 1948, *Terr. Mag.*, **53**, 51.

BATES, D. R., and MCDOWELL, M. R. C., 1957, *J. Atmos. Terr. Phys.*, **10**, 96.
BATES, D. R., and MASSEY, H. S. W., 1947, *Proc. Roy. Soc.* A, **192**, 1.
BECKER, W., 1956, *A.E.U.*, **10**, 207.
BECKER, W., and DIEMINGER, W., 1950 a, *Naturwissenschaften*, **37**, 90 ; 1950 b, *Z. Naturforsch.* A, **5**, 308 ; 1951, *Proc. Mixed Comm. on Ionosph.*, 2nd Meeting, p. 126.
BERG, O. E., KOOMEN, M., MEREDITH, L., and SCOLNIK, R., 1956, *J. Geophys. Res.*, **61**, 302.
BERKNER, L. V., 1939, *Phys. Rev.*, **55**, 536.
BERKNER, L. V., and WELLS, H. W., 1937, *Terr. Mag.*, **42**, 301.
BERNING, W. W., 1954, *Rocket Exploration of the Upper Atmosphere* (London : Pergamon Press), p. 261.
BEST, J. E., FARMER, F. T., and RATCLIFFE, J. A., 1938, *Proc. Roy. Soc.* A, **164**, 96.
BEST, J. E., and RATCLIFFE, J. A., 1938, *Proc. Phys. Soc.*, **50**, 233.
BEYNON, W. J. G., 1955, *Proc. Mixed Comm. on Ionosph.*, 4th Meeting, p. 31.
BEYNON, W. J. G., and BROWN, G. M., 1951, *Proc. Mixed Comm. on Ionosph.*, 2nd Meeting, p. 156 ; 1956, *Nature, Lond.*, **177**, 583 ; 1957, *Observatory*, **77**, 94.
BIBL, K., 1950, *Naturwissenschaften*, **37**, 373 ; 1951, *Ann. Géophys.*, **7**, 208 ; 1953, *Z. Geophysik*, **19**, 136.
BIBL, K., BUSCH, R., RAWER, K., and SUCHY, K., 1955, *J. Atmos. Terr. Phys.*, **6**, 69.
BIONDI, M. A., and BROWN, S. C., 1949, *Phys. Rev.*, **76**, 1697.
BREIT, G., and TUVE, M. A., 1925, *Terr. Mag.*, **30**, 15.
BYRAM, E. T., CHUBB, T. A., and FRIEDMAN, H., 1954, *J. Opt. Soc. Amer.*, **44**, 353 ; 1956, *J. Geophys. Res.*, **61**, 251.
BYRAM, E. T., CHUBB, T. A., FRIEDMAN, H., and GAILAR, N., 1953, *Phys. Rev.*, **92**, 1066.
CHAPMAN, S., 1931, *Proc. Phys. Soc.*, **43**, 26 and 483.
CHUBB, T. A., FRIEDMAN, H., KREPLIN, R. W., and KUPPERIAN, J. E., 1957, *Nature, Lond.*, **179**, 861.
COMPTON, A. H., and ALLISON, S. K., 1954, *X-rays in Theory and Experiment* (New York : Van Nostrand), p. 524.
DELLINGER, J. H., 1935, *Phys. Rev.*, **48**, 705 ; 1937, *Terr. Mag.*, **42**, 49.
DENISSE, J. E., and KUNDU, M. R., 1957, *C.R. Acad. Sci., Paris*, **244**, 45.
DIEMINGER, W., and GEISWEID, K. H., 1950, *J. Atmos. Terr. Phys.*, **1**, 42.
DIEMINGER, W., and HOFFMANN-HEYDEN, A. E., 1952, *Naturwissenschaften*, **4**, 84.
ELWERT, G., 1953, *J. Atmos. Terr. Phys.*, **4**, 68.
FINDLAY, J. W., 1947, *Nature, Lond.*, **159**, 58 ; 1951 a, *J. Atmos. Terr. Phys.*, **1**, 353 ; 1951 b, *Ibid.*, **1**, 367.
FRIEDMAN, H., 1955, *Ann. Géophys.*, **11**, 174.
FRIEDMAN, H., LICHTMAN, S. W., and BYRAM, E. T., 1951, *Phys. Rev.*, **83**, 1025.
GAZZARD, A. D., 1958, *Aust. J. Phys.*, **11**, 272.
GILLILAND, T. R., 1935, *Proc. Inst. Radio Engrs*, **23**, 1076.
GLEDHILL, J. A., and SZENDREI, M. E., 1950, *Proc. Phys. Soc.* B, **63**, 427.
GRACE, C. H., 1951 a, *J. Geophys. Res.*, **56**, 452 ; 1951 b, *Penn. State Uni. Ionosph. Rep.* No. 22.
GREGORY, J. B., 1956, *Aust. J. Phys.*, **9**, 324.
HAGG, E. L., and HANSON, G. H., 1954, *Canad. J. Phys.*, **32**, 790.
HALLIDAY, E. C., 1936, *Proc. Phys. Soc.*, **48**, 421.
HANSON, G. H., HAGG, E. L., and FOWLE, D., 1953, *Defence Res. Telecomm. Estab.* (*Ottawa*), R.P.L. Rep. No. R-2.
HARANG, L., 1937, *Terr. Mag.*, **42**, 55.
HARNISMACHER, E., 1950, *C.R. Acad. Sci., Paris*, **230**, 1301 ; 1955, *Ibid.*, **240**, 553.
HARTREE, D. R., 1929, *Proc. Camb. Phil. Soc.*, **25**, 47.
HAVENS, R. J., FRIEDMAN, H., and HULBURT, E. O., 1955, *Physics of the Ionosphere* (London : Physical Society), p. 237.
HAVENS, R. J., KOLL, R. T., and LAGOW, H. E., 1952, *J. Geophys. Res.*, **57**, 62.
HENDERSON, J. T., 1933, *Canad. J. Phys.*, **8**, 1.
HEPPNER, J. P., and MEREDITH, L. H., 1958, *J. Geophys. Res.*, **63**, 51.
HIGGS, A. J., 1942, *Mon. Not. R. Astr. Soc.*, **102**, 24.

Hok, G., and Dow, W. G., 1954, *Rocket Exploration of the Upper Atmosphere* (London : Pergamon Press), p. 240.
Hollingworth, J., 1935, *Proc. Phys. Soc.*, **47**, 843.
Horowitz, R., and LaGow, H. E., 1957, *J. Geophys. Res.*, **62**, 57.
Houston, R. E., 1957, *Penn. State Uni. Ionosph. Rep.* No. 95 ; 1958, *J. Atmos. Terr. Phys.*, **12**, 225.
Hulburt, E. O., 1939, *Phys. Rev.*, **55**, 639.
Jackson, J. E., 1954, *J. Geophys. Res.*, **59**, 377 ; 1956, *Ibid.*, **61**, 107.
Jackson, J. E., Kane, J. A., and Seddon, J. C., 1956, *J. Geophys. Res.*, **61**, 749.
Jackson, J. E., and Seddon, J. C., 1958, *J. Geophys. Res.*, **63**, 197.
Johnson, C. Y., Purcell, J. D., and Tousey, R., 1954, *Phys. Rev.*, **95**, 621.
Kaiser, T. R., and Evans, S., 1955, *Ann. Géophys.*, **11**, 148.
Kelso, J. M., 1952, *J. Geophys. Res.*, **57**, 357.
King, G. A. M., 1957, *J. Atmos. Terr. Phys.*, **11**, 209.
Kirby, S. S., and Judson, E. B., 1935, *Proc. Inst. Radio Engrs*, **23**, 733.
Lien, J. R., Marcou, R. J., Ulwick, J. C., Aarons, J., and McMorrow, D. R., 1954, *Rocket Exploration of the Upper Atmosphere* (London : Pergamon Press), p. 223.
Lyon, A. J., and Moorat, A. J. G., 1956, *J. Atmos. Terr. Phys.*, **8**, 309.
Mange, P., 1954, *Penn. State Uni. Ionosph. Rep.* No. 64 ; 1955, *Ann. Géophys.*, **11**, 153.
Martyn, D. F., 1948, *Proc. Roy. Soc.* A, **194**, 429 ; 1953, *Phil. Trans. Roy. Soc.*, **246**, 306.
Martyn, D. F., and Green, A. L., 1933, *Nature, Lond.*, **132**, 523.
Matsushita, S., 1953, *Rep. Ionosph. Res. Japan*, **7**, 45.
McNichol, R. W. E., and Gipps, G. de V., 1951, *J. Geophys. Res.*, **56**, 17.
Meadows, E. B., and Townsend, J. W., 1958, *Ann. Géophys.*, **14**, 80.
Menzel, D. H., and Wolbach, J. G., 1955, *Proc. Mixed Comm. on Ionosph.*, 4th Meeting, p. 27.
Minnis, C. M., 1956 a, *J. Atmos. Terr. Phys.*, **9**, 36 ; 1956 b, *Radio Res. Board Rep.* (G.B.), No. IC/34 ; 1957 a, *Observatory*, **77**, 94 ; 1957 b, *J. Atmos. Terr. Phys.*, **10**, 229 ; 1958 a, *Ibid.*, **12**, 272 ; 1958 b, *Ibid.*, **12**, 266.
Mitra, A. P., and Jones, R. E., 1953, *Penn. State Uni. Ionosph. Rep.* No. 44.
Mitra, A. P., and Jones, R. E., 1954, *J. Geophys. Res.*, **59**, 391.
Mitra, S. K., 1952, *The Upper Atmosphere* (Calcutta : Royal Asiatic Soc. of Bengal).
Mögel, H., 1931, *Wireless Engineer*, **8**, 604.
Naismith, R., 1943, *Radio Res. Board Rep.* (G.B.), No. C.75 ; 1954, *J. Atmos. Terr. Phys.*, **5**, 73.
Nicolet, M., 1954, *Penn. State Uni. Ionosph. Rep.* No. 61.
Nicolet, M., and Bossy, L., 1949, *Ann. Géophys.*, **5**, 275.
Nicolet, M., and Mange, P., 1954, *J. Geophys. Res.*, **59**, 15.
Pfister, W., and Ulwick, J. C., 1958, *J. Geophys. Res.*, **63**, 315.
Piddington, J., 1951, *J. Geophys. Res.*, **56**, 409.
Rastogi, R. G., 1954, *Proc. Ind. Acaa Sci.* A, **40**, 158 ; 1956, *J. Atmos. Terr. Phys.*, **9**, 71.
Ratcliffe, J. A., 1956, *Solar Eclipses and the Ionosphere*, Ed. W. J. G. Beynon and G. M. Brown (London : Pergamon Press), p. 1.
Ratcliffe, J. A., and White, E. L. C., 1934, *Proc. Phys. Soc.*, **46**, 107.
Rawer, K., 1953, *Die Ionosphäre* (Groningen : P. Noordhoff ; New York, 1957 : Frederick Ungar Publishing Co. (English translation)) ; 1955, *Proc. Mixed Comm. on Ionosph.*, 4th Meeting, p. 107.
Rawer, K., and Argence, E., 1953, *Ann. Géophys.*, **9**, 1.
Robinson, B. J., 1957, *Thesis*, University of Cambridge.
Saha, A. K., and Ray, S., 1955, *J. Atmos. Terr. Phys.*, **7**, 107.
Sato, T., 1955, *Rep. Ionosph. Res. Japan*, **9**, 205.
Schafer, J. P., and Goodall, W. M., 1933, *Nature, Lond.*, **131**, 804.
Scott, J. C. W., 1952, *J. Geophys. Res.*, **57**, 369.
Seaton, S. L., 1948, *J. Met.*, **5**, 204.
Seddon, J. C., 1954 a, *J. Geophys. Res.*, **59**, 463 ; 1954 b, *Rocket Exploration of the Upper Atmosphere* (London : Pergamon Press), p. 214.

SEDDON, J. C., PICKAR, A. D., and JACKSON, J. E., 1954, *J. Geophys. Res.*, **59**, 513.
SHIMAZAKI, T., 1956, *J. Radio Res. Lab.* (*Japan*), **3**, 17 ; 1957, *Ibid.*, **4**, 37 ; 1958, *Ibid.*, **5**, 35.
SHINN, D. H., and WHALE, H. A., 1952, *J. Atmos. Terr. Phys.*, **2**, 85.
SMITH-ROSE, R. L., and BARFIELD, R. H., 1926, *Proc. Roy. Soc.* A, **110**, 580.
TAUBENHEIM, J., 1957, *J. Atmos. Terr. Phys.*, **11**, 14.
THOMAS, J. A., and SVENSON, A. C., 1955, *Aust. J. Phys.*, **8**, 554.
TOWNSEND, J. W., MEADOWS, E. B., and PRESSLEY, E. C., 1954, *Rocket Exploration of the Upper Atmosphere* (London : Pergamon Press), p. 169.
TREMELLEN, K. W., and COX, J. W., 1947, *J. Instn Elect. Engrs*, Pt. III A, **94**, 200.
WALDMEIER, M., 1944, *Helv. Phys. Acta*, **17**, 168.
WARK, D. Q., and STONE, J. M., 1955, *Nature, Lond.*, **175**, 254.
WATANABE, K., MARMO, F., and PRESSMAN, J., 1955, *J. Geophys. Res.*, **60**, 513.
WAYNICK, A. H., 1957, *Proc. Inst. Radio Engrs*, **45**, 741.
WEEKES, K., 1948, *Thesis, Camb. Uni.*, pp. 65–70, 121 ; 1954, *Quart. J. R. Met. Soc.*, **65**, 324.
WHALE, H. A., 1950, *Thesis, Camb. Uni.*, Ch. 4, 5, 6 ; 1951 a, *J. Atmos. Terr. Phys.*, **1**, 244 ; 1951 b, *Ibid.*, **1**, 233.
WRIGHT, J. W., KNECHT, R. W., and DAVIES, K., 1957, *Annals of the I.G.Y.*, Vol. 3, Part 1. Ed. W. J. G. Beynon and G. M. Brown (London : Pergamon Press).

RADIO-FREQUENCY SPECTROSCOPY OF EXCITED ATOMS

By G. W. SERIES
Clarendon Laboratory, Oxford

CONTENTS

	PAGE
§ 1. Introduction	280
1.1. Precision of measurement	280
1.2. Stimulated transitions and conditions of observation	281
1.3. Types of experiment	282
§ 2. Time-independent theory	285
2.1. Stationary state theory of hyperfine structure	285
2.2. Calculation of the nuclear moments	287
§ 3. Time-dependent theory	289
3.1. Simple model for the radio-frequency transitions	289
3.2. Violation of the simplifying assumptions	291
3.3. Multiplicity of energy levels	292
3.4. Finite life of excited states	294
3.5. Transitions between excited states in the vicinity of other states	297
3.6. Quanta of different kinds	298
§ 4. Double resonance experiments	299
4.1. Resonance radiation	299
4.2. Spatial distribution of resonance radiation: classical model	300
4.3. Intensity and polarization of resonance radiation: quantum theory	301
4.4. Double resonance experiments in magnetic fields	305
4.5. Double resonance experiments in zero magnetic field	311
4.6. Excitation by unpolarized light in zero magnetic field	314
4.7. Double resonance measurements of isotope shifts	316
4.8. Detection by coincidence of photons in cascade	317
§ 5. Excitation by electron impact	317
5.1. Selective excitation of different states within a given term	319
5.2. Selective excitation of different terms	321
§ 6. Atomic beam experiments	323
§ 7. Table of measurements	325
Acknowledgments	327
References	327

Abstract. Radio-frequency and optical spectroscopy are compared. Standard results in the theory of hyperfine structure are quoted. The excitation of atoms and the detection of radio-frequency resonances are discussed quantitatively. Experiments are described in which the techniques of ' double resonance ', electron impact, and atomic beams are used. A table of measurements is given.

§ 1. Introduction

1.1. *Precision of Measurement*

Excited states of atoms have conventionally been studied by optical spectroscopy. In the analysis of a set of close-lying energy levels such as Zeeman or hyperfine structure components interferometric techniques are extremely powerful. The limit to the precision of such work lies not in the apparatus—the Fabry–Pérot interferometer, for example, is capable of whatever

resolving power may be required—but rather in the line width of the radiation under investigation. The most important cause of line broadening in high resolution spectroscopy is the Doppler effect, which leads to half-intensity line widths of order of magnitude 0·02 cm^{-1} (600 Mc/s). (A considerable reduction on this figure is achieved in atomic beam light sources.)

This limitation in the study of close energy levels is not fundamental, but arises merely through the very high frequency of optical radiation: it will be recalled that the Doppler width is proportional to the frequency which is actually observed. A great increase in precision is to be expected if the intervals between energy levels are investigated directly, rather than through the intermediacy of other, very remote, levels.

This report will be concerned with such direct measurements, made in the radio- and microwave-frequency regions. We shall often use the term ' radio-frequency ' alone to cover also the microwave range.

The high precision characteristic of measurements of ground levels is not to be expected in the radio-frequency spectroscopy of excited states because of the finite lifetimes of the atoms. A typical lifetime of 10^{-8} sec implies a broadened energy level and a half-intensity line width of about 30 Mc/s (both levels contribute to the width). The gain in precision of measurement over optical methods is nevertheless very considerable, as these order-of-magnitude figures imply.

In using the term ' excited states ' in this report, we exclude metastable states when the radio-frequency transition is between a pair of metastables. Among the experiments which, for this reason, we ignore are those of Heberle, Reich and Kusch (1956) on the hyperfine structure of the $2^2S_{1/2}$ levels in hydrogen and deuterium, those of Novick and Commins (1958) on the same structure in helium 3, and those of Drake, Lurio, Hughes and White (1958) on the g_J values of the $^3S_1$ level of helium and the $^3P_2$ levels of neon and argon. We include experiments in which only one of the participating states is metastable.

1.2. *Stimulated Transitions and Conditions of Observation*

The very large factor ($\sim 10^6$) between typical optical and radio frequencies implies a factor of the order of 10^{18} in the ratio of the probabilities of spontaneous transitions in the two regions. Thus, while optical spectroscopy relies on the spontaneous decay of excited states, radio-frequency spectroscopy must concern itself with stimulated transitions. Electric dipole and magnetic dipole transitions have been investigated.

The method of detection of transitions in most of the experiments we shall describe requires, in consequence of the equality of transition probabilities forward and backward between a given pair of states, the prior establishment of a difference of population between them. The various methods by which this is accomplished are discussed below.

A further condition for the observation of resonances is that the radio-frequency field must be sufficiently strong to induce transitions before the excited atoms decay. This implies that the matrix elements of $\mathbf{p}.\mathbf{E}$ ($\boldsymbol{\mu}.\mathbf{H}$) between the given pair of states be of the order of $\hbar/\tau$, where $\mathbf{P}$ ($\boldsymbol{\mu}$) is the electric (magnetic) dipole moment, τ is the lifetime of the initially excited state, and $\mathbf{E}$ ($\mathbf{H}$) is the requisite field strength.

An important difference between electric and magnetic dipole transitions is that the former are much stronger. An estimate of the order of magnitude is obtained by comparing the interaction energies of the two types of dipole, of atomic magnitudes, in electric and magnetic fields of comparable strength. The transition probabilities are in the ratio

$$\frac{\mathbf{p}.\mathbf{E}}{\boldsymbol{\mu}.\mathbf{H}} = \frac{ea_0}{e\hbar/2mc} = 2/\alpha \simeq 300$$

where we have used e.s.u. for the electric dipole and e.m.u. for the magnetic, and set the magnitudes of the fields equal in e.s.u. and e.m.u. respectively; a_0 is the Bohr radius and α the fine structure constant.

1.3. *Types of Experiment*

We shall classify the experiments on excited states by the means which they employ to achieve differences in population, and by their methods of detection of radio-frequency resonance.

1.3.1. *The attainment of differences in population.*

Our starting point will be an assembly of atoms in statistical equilibrium in the ground states. We ignore population differences, if any, which are a consequence of the Boltzmann distribution. Convenient means of excitation are by optical resonance radiation and by electron impact.

Isotropic, unpolarized resonance radiation, uniform in intensity over the region of absorption, results in populations which are in statistical equilibrium, and are therefore of no interest to us. Similarly, the excitation of a set of close-lying atomic energy levels by electrons in a gas discharge is generally characterized by statistical equilibrium at some (high) electron temperature.

Population differences may be induced by resonance radiation which is either non-isotropic, or polarized, or restricted in frequency range, or by any combination of these. Similarly, population differences may be induced by controlled electron bombardment, or in gas discharges in which for some reason statistical equilibrium is not maintained. § 4 of this Report is concerned with excitation by resonance radiation, § 5 with excitation by electron impact.

(i) *Excitation by resonance radiation.*

Polarized resonance radiation was used by Brossel and Bitter (1952) in their pioneer ' double resonance ' experiment which is described in § 4.4.1. Readers who are unfamiliar with the subject may find it helpful to refer now to this concrete example. The later sub-sections describe the developments which have taken place in this type of experiment. Measurements of Zeeman splittings, hyperfine structures, isotope shifts and lifetimes of excited states have been made.

Unpolarized resonance radiation was used by Rabi and his collaborators in atomic beam experiments designed to measure hyperfine structures in excited states of the alkali metals. The illuminated atoms were not in statistical equilibrium. This work is described in § 6.

(ii) *Excitation by electron bombardment.*

Controlled electron bombardment was used in the first successful radio-frequency experiment on excited states, that of Lamb and Retherford on the $2^2S_{1/2}$–$2^2P_{1/2}$ interval in atomic hydrogen. We shall not describe this celebrated experiment here, since it has already been described in these Reports (Lamb 1951). Details of the experiment and its interpretation have been published in a set of six papers (Lamb 1952, Lamb and Retherford 1950, 1951, 1952, Dayhoff, Triebwasser and Lamb 1953, Triebwasser, Dayhoff and Lamb 1953). Lamb's 1951 Report describes also the first version of a similar experiment on the analogous level in He^+. We shall describe later versions of this experiment in § 5.2, which also treats other experiments on hydrogen and helium, mainly by Lamb and his collaborators.

The object of these experiments was to make precise measurements of structures in atoms simple enough to be amenable to calculations of very high accuracy. 'Lamb shifts' were discovered and measured in the spectra of hydrogen and ionized helium. Accurate values were obtained for the intervals between triplet levels in neutral helium. In this report we shall not treat the theoretical aspects of the work. A non-specialist account confined to one-electron spectra has recently been published (Series 1957).

Controlled electron bombardment has been used also in experiments which developed naturally from the double resonance experiments. These are described in § 5.1. The technique holds promise of being more versatile than excitation by resonance radiation in that a greater variety of excited states is accessible.

A number of interesting variants on these two methods of excitation have been tried, to which reference will be made in later sections.

1.3.2. *Methods of detection.*

We first consider, quite generally, two different methods for the detection of transitions which have been stimulated by radio-frequency fields:

(*a*) Measurements on the radiation field itself. This is the technique used in nuclear and paramagnetic resonance experiments in which one looks for changes of energy in the stimulating electrical circuit, at the rate of one radio-frequency quantum for each atomic transition. The same limiting sensitivity is obtained by observing changes of phase in the electrical circuit.

(*b*) Observation of some secondary change consequent upon the radio-frequency transition. Since an amount of energy much larger than a radio-frequency quantum may be available per atomic transition, this is potentially a much more sensitive technique than the former.

We shall be exclusively concerned with the detection of resonances by secondary effects. Self-evidently, one might make use of the optical (or other short wavelength) radiation emitted in the decay of the excited state. The major part of this report deals with the detection of radio-frequency resonances through the optical radiation. Other methods of detection are used in the atomic beam experiments.

(i) *Detection by changes in intensity or polarization of the emitted optical radiation.*

The optical radiation with which one is concerned is electric dipole. If, therefore, the radio-frequency transition changes the electric dipole matrix

element between the excited and ground states, then the transition can be detected by a change in the total intensity of light radiated in decay. It is clear that electric dipole radio-frequency transitions will have the required property. On this basis radio-frequency resonances were detected in the experiment on He^+ of Lamb and Skinner (1950) and its recent refinements, and in related experiments on one-electron spectra (§ 5.2).

The change in total intensity is generally of less interest than the change in intensity in some particular direction. Magnetic dipole transitions can bring about such changes, though they do not change the total intensity of the light (unless the atomic spin–orbit coupling is strong). Their effect is to re-orient the atom, and thus to alter the spatial distribution of its (anisotropic) radiation.

Magnetic dipole transitions are formally represented as causing changes in the polarization of the emitted light: indeed, more detailed information about the excited states is sometimes obtainable by examination of the polarization, as well as of the intensity changes. It is, however, worth noticing that magnetic dipole transitions in excited states can often be detected with no polarizing equipment at all.

The experiments of Lamb and his colleagues on the (two-electron) spectrum of neutral helium (§ 5.1.3) make use of changes in intensity of the optical radiation in a given direction to detect magnetic dipole transitions.

In ' double resonance ' experiments (§ 4) changes in the intensity or polarization of the fluorescent radiation are used to detect magnetic dipole transitions.

(ii) *Detection by other changes in the optical radiation.*

We shall mention three further possibilities for detection in double resonance experiments which, so far as we are aware, have not yet been exploited in experiments on excited states. They are the measurement of optical dispersion, of the Faraday effect, and of amplitude modulation of the fluorescent light.

Dispersion of optical radiation arises through interference between the primary beam and light scattered in the forward direction. The effects are large near a resonance line. Radio-frequency resonance between excited states, since it causes a re-distribution of the scattered light, should induce a change in refractive index of the scattering vapour.

The Faraday effect is proportional to the difference in refractive indices for σ^+ and σ^- light. The amount of the Faraday rotation should therefore change when the refractive indices change at radio-frequency resonance. An effect of this sort has been observed to accompany paramagnetic resonance in the ground state of a Cr^{3+} ion in a crystal (Wesemeyer and Daniels 1958). Differences in population were in this case obtained by cooling the crystal in liquid helium.

The third possibility is a consequence of the mixing of states which radio-frequency resonance brings about. Since, under the influence of the radio-frequency field, the state of an excited atom becomes a mixture of the original eigenstates, there will exist definite phase relationships between the scattered light from different levels. The frequencies of these components will differ by the radio frequency itself, so that interference between them will result in modulation at the radio frequency of the amplitude of the scattered light. Experiments

on these lines are in progress. The effect is related to a phenomenon predicted by Dehmelt (1957) and observed by Bell and Bloom (1957) in the ground state of sodium.

(iii) *Detection by other secondary effects.*

In atomic beam experiments radio-frequency transitions are detected by changes in the number of atoms which reach a detector. Types of detector are various: in Lamb's experiments on the $n = 2$ levels in hydrogen the detector consisted of a cold tungsten plate with the property of liberating electrons when struck by metastable hydrogen atoms; in Rabi's experiments on excited levels of alkali metals the detector consisted of a hot tungsten wire which stripped electrons off the alkali atoms and liberated them as ions. In both cases the radio-frequency transition is followed by some very much more energetic change in the detector.

As we have already remarked, we shall not describe Lamb's experiment in this report. We mention it in this introductory section for the sake of completeness. Rabi's experiments are described in § 6.

§ 2. Time-independent Theory

The small energy of radio-frequency quanta limits the application of these techniques to the measurement of hyperfine intervals, apart from a few important special cases in which multiplet intervals are abnormally small, in particular, in the spectra of hydrogen, helium and ionized helium.

For the sake of convenient reference we shall give a few standard results in the theory of hyperfine structure in atomic spectra. Further details may be found in the book by Kopfermann (1958). We shall not discuss the theory of the special cases of multiplet structure.

We shall limit ourselves in § 2 to stationary state (time-independent) theory. In § 3 we discuss transition processes (time-dependent).

2.1. *Stationary State Theory of Hyperfine Structure*

The following Hamiltonian function is generally used to represent the magnetic dipole and electric quadrupole interactions in an atom under a uniform external magnetic field $\mathbf{H}$:

$$\mathscr{H} = A\mathbf{I}.\mathbf{J} + \frac{B}{2I(2I-1)J(2J-1)}[3(\mathbf{I}.\mathbf{J})^2 + \tfrac{3}{2}(\mathbf{I}.\mathbf{J}) - I(I+1)J(J+1)] + g_J\beta\mathbf{J}.\mathbf{H} - g_I\frac{m}{M}\beta\mathbf{I}.\mathbf{H}. \qquad \ldots\ldots(1)$$

In this equation $\mathbf{I}$ and $\mathbf{J}$ are operators, having eigenvalues $[I(I+1)]^{1/2}$ and $[J(J+1)]^{1/2}$ respectively in units $\hbar$, which represent the nuclear and electronic angular momenta. A and B are constants proportional to the nuclear magnetic dipole and electric quadrupole moments, respectively. (The term in B does not apply unless both J and I are greater than $\frac{1}{2}$); g_J and g_I are the electronic and nuclear g-factors, β the Bohr magneton, and m/M the electron/proton mass ratio. The positive sign of the term $g_J\beta\mathbf{J}.\mathbf{H}$ comes from the negative sign of the interaction energy, $-\mu_J.\mathbf{H}$ coupled with the substitution $\mu_J = -g_J\beta\mathbf{J}$ which takes

account of the negative charge on the electron. On the other hand, since the nucleus is positively charged we define g_I through $\mu_I = g_I \beta \mathbf{I}m/M$, so that the last term in (1) has a negative sign.

From (1) the hyperfine energy can be deduced in terms of the constants A and B, which are then determined by comparison with experiment. Values for the nuclear magnetic dipole and electric quadrupole moments are subsequently obtained under various simplifying assumptions.

2.1.1. *Zero magnetic field.*

We choose a representation in which F ($\mathbf{F} = \mathbf{I}+\mathbf{J}$) is a good quantum number. The energy of the level F is then found to be

$$W(F) = \tfrac{1}{2}AC + \frac{B}{I(2I-1)J(2J-1)}[\tfrac{3}{8}C(C+1) - \tfrac{1}{2}I(I+1)J(J+1)], \quad \ldots\ldots(2)$$

where $C \equiv F(F+1) - I(I+1) - J(J+1)$.

As an example of the working out of this formula, we quote the hyperfine structure intervals for a nucleus of spin $I = \frac{3}{2}$ in interaction with electronic spins $J = \frac{1}{2}$ and $J = \frac{3}{2}$:

(i) $J = \frac{1}{2}$ $\quad W_2 - W_1 = 2A$

(ii) $J = \frac{3}{2}$ $\quad W_3 - W_2 = 3A + B$

$\qquad\qquad\quad W_2 - W_1 = 2A - B$

$\qquad\qquad\quad W_1 - W_0 = A - B.$

2.1.2. *Weak magnetic field.*

In a magnetic field the terms $g_J \beta \mathbf{J}.\mathbf{H}$ and $g_I \beta \mathbf{I}.\mathbf{H}m/M$ introduce energy additional to the value of $W(F)$ given by (2). Under the approximation $g_J \beta H \ll A, B$ the additional energy is

$$\Delta W(F, m_F) = g_F \beta \mathbf{F}.\mathbf{H} = g_F \beta m_F H \quad \ldots\ldots(3)$$

where

$$g_F = g_J\left[\frac{F(F+1)+J(J+1)-I(I+1)}{2F(F+1)}\right] - g_I \frac{m}{M}\left[\frac{F(F+1)+I(I+1)-J(J+1)}{2F(F+1)}\right]$$

and m_F takes the $2F+1$ values $F, F-1 \ldots -F$.

The second term in the equation for g_F is three orders of magnitude smaller than the first. Its effect on ΔW is generally small compared with the natural width of the energy levels of excited states.

2.1.3. *Strong magnetic field.*

Under the approximation $g_J \beta H$ $(= G) \gg A, B$ we may again give a simple expression for the energy in the representation in which m_J and m_I are good quantum numbers:

$$W(m_J, m_I) = G\left(m_J - \frac{g_I}{g_J}\frac{m}{M} m_I\right) + A m_J m_I + \frac{B}{4I(2I-1)J(2J-1)}[3m_J^2 - J(J+1)][3m_I^2 - I(I+1)]. \quad \ldots\ldots(4)$$

Again, the term in $g_I m/M$ may often be neglected.

The order of magnitude of the most important correction terms, which may be evaluated by standard perturbation theory, is $A^2, B^2/G$.

2.1.4. *Magnetic fields of intermediate strength.*

The secular equation whose solution yields the general result may always be arranged so that it factorizes into a set of equations whose first member is linear, and the others of progressively higher order in the energy. If either I or J does not exceed $\frac{1}{2}$, no equation of order higher than quadratic appears. The solution for $J = \frac{1}{2}$, I arbitrary, is

$$W(\pm\tfrac{1}{2}, m_I) = -\tfrac{1}{4}A - (m_I \pm \tfrac{1}{2})g_I \frac{m}{M}\beta H \pm \tfrac{1}{2}[A^2(I+\tfrac{1}{2}) + 2AG'(m_I \pm \tfrac{1}{2}) + G'^2]^{1/2} \quad \ldots\ldots(5)$$

where

$$G' = \left(g_J + g_I \frac{m}{M}\right)\beta H.$$

2.2. *Calculation of the Nuclear Moments*

The constants A and B express the interaction of the nucleus with all the electrons. Under an approximation which is very good for the magnetic dipole, but rather crude for the electric quadrupole interaction, the contributions from closed shells are generally ignored in the preliminary calculation. More refined calculations have been made by Sternheimer (1954).

Individual electrons outside closed shells each contribute to A and B in a way which depends on their coupling. Contributions from individual electrons are vectorially additive. Details, for certain configurations of two electrons, are given by Kopfermann (1958). We shall quote the results for a one-electron system specified by orbital and total angular momentum quantum numbers l and j. Following Kopfermann, we use a, b, to denote the interaction constants for individual electrons.

2.2.1. *The nuclear gyromagnetic ratio g_I.*

For an s-electron ($l = 0$, $j = \frac{1}{2}$) one finds

$$g_I = a_{1/2}\frac{3n_0^3}{8R_\infty \alpha^2 Z Z_o^2 (1 - d\sigma/dn)} \quad \ldots\ldots(6)$$

apart from certain correction terms of order unity. Z is the nuclear charge, Z_o the 'outer' nuclear charge (1 for a neutral atom, 2 for single ionization, etc.), R_∞ the Rydberg constant, α the fine structure constant, and n_0 the effective principal quantum number defined by $(T/RZ_o^2)^{1/2}$, where T is the term value (measured from the ionization limit). σ is the quantum defect $n - n_0$.

For $l \neq 0$, both a and b are proportional to the mean value of $1/r^3$, where r is the electron–nuclear distance. Uncertainty in the value of this quantity can be eliminated by the introduction of δ, the fine structure doublet splitting, since

$\overline{1/r^3}$ also enters into a calculation of this interval. We then find, for either of the factors $a_j = a_{l\pm\frac{1}{2}}$,

$$g_I = a_j \frac{Z_i(l+\frac{1}{2})j(j+1)}{\delta l(l+1)} \quad (l \neq 0), \qquad \ldots\ldots(7)$$

apart from corrections of order unity as before. Z_i is the 'inner' nuclear charge whose value is approximately the true nuclear charge Z diminished by 4 for a p-electron and by 11 for a d-electron. A more exact definition is contained in Casimir's expression (1936) for the doublet fine structure,

$$\delta = R_\infty \alpha^2 \frac{dn_0}{dn} \frac{Z_o^2 Z_i^2}{n_0^3 l(l+1)} \mathscr{R} \qquad \ldots\ldots(8)$$

in which n is the true principal quantum number and $\mathscr{R}$ a relativistic correction.

Corrections to these expressions for g_I arise when one takes account of the finite size of the nucleus and the distribution of magnetic moment within it, and also when one introduces relativistic wave functions into the theory. These corrections also are discussed by Kopfermann.

Equation (7) implies that a simple relation exists between the two a factors. It is

$$a_{l+\frac{1}{2}} = \left(\frac{2l-1}{2l+3}\right) a_{l-\frac{1}{2}}, \qquad \ldots\ldots(9)$$

apart from relativistic corrections. Deviations from this relation of about 14% have been found in rubidium, for example (Senitzky and Rabi 1956). They can be understood on the basis that the allocation of a single quantum state (n, l) to each electron is only an approximation. A consequence of allowing that a small fraction of some other configuration may be mixed with a particular configuration is to alter the a factors: the fractional change in $a_{l+\frac{1}{2}}$ is not necessarily the same as that in $a_{l-\frac{1}{2}}$. On these lines Phillips (1956) has given a reasonable explanation of the discrepancy in rubidium.

Deviations from (9) are also to be expected when the uncertainty of measurement is less than a^2/δ, for when this quantity is significant the perturbation between the doublet levels $j = l \pm \frac{1}{2}$ is significant. An especially interesting case arises in *Li* II where a and δ are the same order of magnitude (Güttinger and Pauli 1931, Gray 1933).

The case of small perturbations has been discussed by Clendenin (1954).

2.2.2. *The nuclear electric quadrupole moment Q.*

In the relation between b and Q it is again desirable to eliminate $\overline{1/r^3}$ in favour of some directly measurable quantity. This can be done either by introducing one of the a factors ($a_{l\pm\frac{1}{2}}$), together with g_I (determined by some other method), or by introducing, as before, δ and Z_i.

Then we find, by the one substitution

$$Q = \frac{b}{a}\left(\frac{\beta}{e}\right)^2 g_I \frac{m}{M} \frac{4l(l+1)}{j(2j-1)}, \qquad \ldots\ldots(10)$$

and by the other,

$$Q = \frac{b}{\delta}\left(\frac{\beta}{e}\right)^2 Z_i \frac{(2l+1)(2j+2)}{2j-1}, \qquad \ldots\ldots(11)$$

where we have omitted relativistic corrections of order unity. β is the Bohr magneton and e the electronic charge.

Taking (10) and (9) we have a third relation between b and Q.

These three relations do not always give consistent results. Acute discrepancies in gallium and other elements have been discussed by Koster (1952), and may be understood on the basis of configuration mixing. The relation (11) appears to be the most reliable.

To the value of Q computed as above from interactions with the outer electrons one must make a correction in consequence of the polarization of the inner core of electrons by the nucleus, by which the electric field gradient at the nucleus is altered. The change is proportional to Q. The necessary corrections may be 10% or more. Sternheimer (1954) has made detailed calculations of this effect. Polarization of the core affects also the relation between a and g_I.

§ 3. Time-dependent Theory

The interaction between an atom of magnetic moment $\boldsymbol{\mu}$ and a magnetic field of amplitude $\mathbf{H}_1$ which rotates in the xy plane with angular frequency ω is expressed by the addition to the Hamiltonian (1) of a term

$$\begin{aligned}\mathscr{H}(t) = -\boldsymbol{\mu}.\mathbf{H}_1(t) &= -\boldsymbol{\mu}.(H_1\cos\omega t\,\mathbf{i} - H_1\sin\omega t\,\mathbf{j}) \\ &= -\tfrac{1}{2}H_1[\mu_+ e^{i\omega t} + \mu_- e^{-i\omega t}], \qquad \ldots\ldots(12)\end{aligned}$$

where $\mu_\pm = \mu_x \pm i\mu_y$.

The magnetic interaction is taken as an example: the argument also covers electrical interactions, in which case we shall use the symbols $\mathbf{p}$ and $\mathbf{E}_1$.

There will generally exist in the atom energy eigenstates between which the matrix elements of $\boldsymbol{\mu}.\mathbf{H}_1$ or $\mathbf{p}.\mathbf{E}_1$ are not zero. The time-dependent perturbation $\mathscr{H}(t)$ will induce transitions to and fro between such states. We are interested in the transition probability as a function of the frequency and of the matrix elements of the perturbation; we may then derive the form of the resonance curve.

The analysis of this problem expressed in its simplest term is well known. It is given, for example, by Ramsey (1956), where references to the literature may be found. The standard solution, which we quote in § 3.1, is obtained under certain assumptions which are not rigorously valid in the experiments we are concerned with. From this solution, nevertheless, it is easy to obtain the equation of the resonance curve for magnetic dipole transitions between an isolated pair of excited states (§ 3.4.1). Experimental observations are normally interpreted with reference to this equation. The case of electric dipole transitions is treated rather differently (§ 3.4.3).

We normally face more complicated situations than an isolated pair of excited states. In the later sub-sections we shall discuss the attempts which have been made to meet these actual situations.

3.1. *Simple Model for the Radio-frequency Transitions*

We consider two energy eigenstates of an atomic system, p and q, of energies W_p and W_q. We define $\omega_0 = (W_q - W_p)/\hbar$. It is to be noticed that in order to specify p and q we must select an axis of quantization. For the purposes of the

following analysis, the axis of quantization will be the z axis of the rotating field, though it need not necessarily be so.

We now consider the effect on the atomic system of the perturbation $-\boldsymbol{\mu}.\mathbf{H}_1(t)$, $(-\mathbf{p}.\mathbf{E}_1(t))$. We shall need the elements of the perturbation matrix

$$\langle p|\mathscr{H}'(t)|q\rangle = -\int \psi_p^*(\tfrac{1}{2}H_1\mu_+ e^{i\omega t})\psi_q\, d\tau = -\hbar\tfrac{1}{2}b_{pq}e^{i\omega t},$$

$$\langle q|\mathscr{H}'(t)|p\rangle = -\int \psi_q^*(\tfrac{1}{2}H_1\mu_- e^{-i\omega t})\psi_p\, d\tau = -\hbar\tfrac{1}{2}b_{qp}e^{-i\omega t}. \qquad \ldots\ldots(13)$$

The magnetic moment $\boldsymbol{\mu}=\boldsymbol{\mu}_L+\boldsymbol{\mu}_S+\boldsymbol{\mu}_I$. We shall ignore the nuclear component and write $\boldsymbol{\mu}\simeq\boldsymbol{\mu}_L+\boldsymbol{\mu}_S$, whose matrix elements are the same as those of $\boldsymbol{\mu}_J=-g_J\beta\mathbf{J}$ between states in which J does not change, that is, for all transitions between Zeeman and hyperfine-structure components. Matrix elements of $\boldsymbol{\mu}_L+\boldsymbol{\mu}_S$ between states which differ in J are given by Lamb (1957). If J does not change, the quantities b_{pq}, b_{qp} are equal to the matrix elements of $J_\pm=(J_x\pm iJ_y)$ multiplied by $g_J\beta H_1/\hbar=\gamma_J H_1$.

In the solution of the problem the following restrictive assumptions are now made: (i) neither p nor q interacts with any other state, (ii) the diagonal elements b_{pp}, b_{qq} are zero. (This is automatically satisfied in our case, since $H_1(t)$ has no component along the z-axis.) An expression can now be deduced rigorously for the probability $P_{pq}(t)$ that an atom, initially in state p, will be in state q after time t. We find

$$P_{pq}(t)=\frac{|b_{pq}|^2}{(\omega_0-\omega)^2+|b_{pq}|^2}\sin^2\{\tfrac{1}{2}[(\omega_0-\omega)^2+|b_{pq}|^2]^{1/2}t\}. \qquad \ldots\ldots(14)$$

(In the similar, but not identical expression first given by Rabi (1937), the states p and q are defined with reference to a rotating axis of quantization.)

The expression (14) forms the basis of all radio-frequency resonance curves. It will be noticed that the first factor on the right-hand side describes a curve with a maximum at $\omega=\omega_0$ and a width at half intensity of $2|b_{pq}|$. The expression for $P_{pq}(t)$ is still a good approximation near resonance when assumption (ii) is violated, provided that $b_{pp}, b_{qq}\ll\hbar\omega_0$.

From (14) we readily deduce the relative probabilities of radio-frequency transitions. For at resonance ($\omega=\omega_0$), and in a time such that the $\sin^2(\ldots)t$ term is still small, we have

$$P_{pq}(t)=\sin^2\tfrac{1}{2}\,b_{pq}|t\to(\tfrac{1}{2}|b_{pq}|t)^2 \qquad \ldots\ldots(15)$$

so that transition probabilities are proportional to the squares of the matrix elements of the perturbation. The probabilities of different atomic transitions in a given radio-frequency field $H_1(t)$ are thus proportional to the squares of the matrix elements of $J_\pm$. All non-vanishing matrix elements of J in the representations (F, m_F) and (J, m_J) are given below. The formulae may be found in Condon and Shortley

(1951), pp. 63, 65, 66.

$$\langle J, m_J | J_z | J, m_J \rangle = m_J$$
$$\langle J, m_J \pm 1 | J_\pm | J, m_J \rangle = [(J \mp m_J)(J \pm m_J + 1)]^{1/2}, \quad \ldots\ldots(16a)$$

$$\langle F, m_F | J_z | F, m_F \rangle = A m_F$$
$$\langle F, m_F | J_\pm | F, m_F \mp 1 \rangle = A[(F \pm m_F)(F \mp m_F + 1)]^{1/2}$$
$$A = \frac{J(J+1) - I(I+1) + F(F+1)}{2F(F+1)}, \quad \ldots\ldots(16b)$$

$$\langle F, m_F | J_z | F+1, m_F \rangle = B[(F+1)^2 - m_F^2]^{1/2}$$
$$\langle F, m_F | J_\pm | F+1, m_F \mp 1 \rangle = \pm B[(F \mp m_F + 1)(F \mp m_F + 2)]^{1/2}$$
$$B = \left[\frac{(F+1-J+I)(F+1+J-I)(J+I+2+F)(J+I-F)}{4(F+1)^2(2F+1)(2F+3)}\right]^{1/2}, \quad \ldots\ldots(16c)$$

$$\langle F, m_F | J_z | F-1, m_F \rangle = C[F^2 - m_F^2]^{1/2}$$
$$\langle F, m_F | J_\pm | F-1, m_F \mp 1 \rangle = \mp C[(F \pm m_F)(F \pm m_F - 1)]^{1/2}$$
$$C = \left[\frac{(F-J+I)(F+J-I)(J+I+1+F)(J+I+1-F)}{4F^2(2F-1)(2F+1)}\right]^{1/2}. \quad \ldots\ldots(16d)$$

Certain matrix elements for the case of intermediate field strengths are given by Torrey (1941).

It will be noticed that J_z has certain non-vanishing matrix elements. If, therefore, there is a field component $H_{1z}(t)$ along the axis of quantization, the diagonal elements of the perturbation, b_{pp}, b_{qq} do not vanish.

The existence of matrix elements of J is expressed by the following selection rules for magnetic dipole transitions between states of the same J:

$\Delta F = 0, \pm 1$; $\Delta m_F = 0, \pm 1$ for the (F, m_F) representation;

$\Delta m_J = 0, \pm 1$; $\Delta m_I = 0$ for the (m_J, m_I) representation.

Transitions $\Delta J = \pm 1$ are also allowed by magnetic dipole interaction.

The transitions $\Delta m_I = 0, \pm 1$; $\Delta m_J = 0$, which are allowed if the nuclear magnetic moment is not neglected, are weaker by the factor $(\mu_I/\mu_J)^2$.

The matrix elements of the electric dipole moment depend on an evaluation of the radial integrals $\int R_p^* r R_q \, dr$ which, for hydrogen wave functions, are tabulated by Condon and Shortley (1951), p. 133. The selection rules are identical with the well-known rules for spontaneous electric dipole transitions.

3.2. *Violation of the Simplifying Assumptions*

In experiments on excited states the assumptions under which equation (14) is derived may be violated in a variety of ways. First, in virtue of the spontaneous emission of radiation, the excited states are connected with remote states. We shall see in § 3.4 how the introduction of decay constants allows a satisfactory treatment of this situation, and generally leads to familiar Lorentz-type expressions for the resonance curve.

The restrictive conditions are again violated if the excited states p and q are connected by the perturbation with other excited states in the vicinity. If the frequency separations are exactly equal a rigorous generalization of equation (14) may be given (§ 3.3.1). Approximate solutions (§ 3.3.2) have been found for the general case of unequal intervals. The application of these results to magnetic dipole transitions between excited states is straightforward (§ 3.4.1). Certain particular cases where the intervals are unequal (or the states are perturbed by other fields) have been examined in detail (§ 3.5).

Violation of the condition that the diagonal elements of the perturbation matrix be small can usually be avoided by arranging that the radio-frequency field be at right angles to the axis of quantization, but it is sometimes of interest to use other geometries. In the exploration of $\Delta m_F = 0$ transitions between weak field hyperfine levels, for example, the condition can easily be violated. No theoretical analysis of this case has come to our notice.

In § 3.6 we shall mention the situation in which the radio-frequency field is composed of several types of quanta. The most common case is met when an oscillating rather than a rotating field is used to induce transitions, which is the usual practice. An oscillating field comprises a mixture of right- and left-handed circularly polarized quanta of the same frequency. The superposition of fields of different frequency is also of interest.

3.3. *Multiplicity of Energy Levels*

3.3.1. *Equally spaced energy levels.*

The groups of states (F, m_F) (or (J, m_J)) include more than two if F (or J) is greater than $\frac{1}{2}$. In sufficiently weak magnetic fields the energy levels are equally spaced. Through the interaction of several radio-frequency quanta transitions may be induced between any pair of the group. Majorana's well-known solution for the probability of a transition (taken from Ramsey 1956, p. 154) is

$$P(F,m,m',t) = (F-m)!\,(F+m)!\,(F-m')!\,(F+m')!\,(\sin\tfrac{1}{2}\alpha)^{4F} \times\left[\sum_r \frac{(-1)^r(\cot\tfrac{1}{2}\alpha)^{m+m'+2r}}{(F-m-r)!\,(F-m'-r)!\,(m+m'+r)!\,r!}\right] \qquad \ldots\ldots(17)$$

in which the summation index r may assume all values for which no argument of a factorial becomes negative, and in which α is defined by the equation

$$\sin^2\tfrac{1}{2}\alpha = P(\tfrac{1}{2},\tfrac{1}{2},-\tfrac{1}{2},t),$$

i.e. the function P_{pq} of equation (14) with

$$\omega_0 = (W_q - W_p)/\hbar = (g_F\beta H_0)/\hbar = \gamma_F H_0$$

and

$$|b_{pq}|^2 = (\gamma_F H_1)^2.$$

It is to be noticed that the probability of a single quantum transition $(m-m' = \pm 1)$ is dependent on whether there exist other values of m within the group. Multiple quantum transitions are not simply represented as the product of successive single transitions. For small values of $\gamma H_1 t$ the probability of a k-quantum transition is proportional to $(\gamma H_1)^{2k}$.

3.3.2. *Unequally spaced energy levels.*

We consider groups of energy levels whose unequal spacing may arise through the onset of I, J (or L, S) decoupling, or because they belong to different values of F in hyperfine structures, or for some other reason. The common feature is that the different members of the group are interconnected by the perturbing field. Transitions $m \to m \pm k$, (k integral), may be induced through the intervention of k equal quanta. Conservation of angular momentum between the radiation field and the atom is thus assured. Conservation of energy requires that

$$W_q - W_p = k\hbar\omega_0.$$

When $k > 1$, the process is described as the simultaneous interaction of k quanta, rather than as k successive interactions. A number of authors (Besset, Horowitz, Messiah and Winter 1954, Winter 1955, Salwen 1955, Hack 1956) have treated the problem theoretically.

The main features disclosed by these analyses are :

(i) The transition probability is generally given by expressions similar to P_{pq} (equation (14)).

(ii) The resonance line width is of the order δ/k, where δ is the width appropriate to single-quantum transitions. It is easy to understand this ; for if k quanta act simultaneously, we shall find $(e^{i\omega t})^k$ in the perturbation operator, and the resonance denominator $[(k\omega - \omega_{pq})^2 + |b_{pq}|^2 + (1/T)^2]^{-1}$ (see eqn (18)) will be proportional to $[(\omega - \omega_0)^2 + |b_{pq}|^2/k^2 + 1/T^2k^2]^{-1}$, where $\omega_0 = \omega_{pq}/k = (W_q - W_p)/k\hbar$.

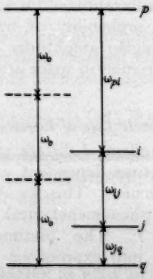


Figure 1. Multiple quantum resonances between unequally spaced levels.

The half-intensity width therefore tends to $2/Tk$ when b_{pq} is small, instead of to $2/T$.

(iii) The resonant frequency ω^* differs from $\omega_0 = (W_p - W_q)/k\hbar$. Analytical expressions (Salwen 1955, p. 1284) for the amount of the shift may be understood on the basis of time-independent perturbation theory in which the effect of the radio-frequency field (to first order) has been taken into account by shifting all the levels by the amount $m\omega_0$ and by allowing perturbation between levels which differ by 1 in m. Intervals between such levels are reduced by the amount ω_0.

Thus, for example, if p, i, j, q is a set of states whose m-values decrease successively by 1, their separations, for the present purpose, become

$$\Delta_{pi} = \omega_{pi} - \omega_0, \quad \Delta_{ij} = \omega_{ij} - \omega_0, \quad \Delta_{jq} = \omega_{jq} - \omega_0.$$

Then if $\frac{1}{2}b_{pi}$ is the matrix element of the perturbation between states p and i, a second order perturbation treatment introduces an upward shift of p by $|b_{pi}|^2/4\Delta_{pi}$ and a downward shift of q by $|b_{jq}|^2/4\Delta_{jq}$. (Since Δ_{jq} is numerically negative in the figure, q is displaced upwards.)

Subject to the usual approximations of perturbation theory ($b \ll \Delta$), we find the frequency shift

$$\omega^* - \omega_0 = \frac{1}{k}\left[\frac{|b_{pi}|^2}{4\Delta_{pi}} + \frac{|b_{jq}|^2}{4\Delta_{jq}}\right]$$

to which the general formulae reduce in our example.

(iv) The intensity of a k-quantum transition under the above approximation is, in our example, proportional to $[|b_{pi}|\,|b_{ij}|\,|b_{jq}|/(\omega_0 - \omega_{qi})(\omega_0 - \omega_{qj})]^2$: the significant feature of this expression is the proportionality (through the b) to the $2k$th power of the radio-frequency field. This is true also in the general case. The higher quantum transitions show a strong dependence on the perturbing field strength.

Multiple quantum transitions have been observed on several occasions in experiments on excited states (Althoff 1955, Ritter and Series 1957). Kusch (1956) undertook a systematic study of these transitions in the ground state of ^{39}K, and found the theory well substantiated for amplitudes of the radio-frequency field which gave a maximum probability of transition. The discrepancies in line width which he found at higher amplitudes can probably be explained by a more detailed analysis of the variation in space of the radio-frequency field (Barrat and Winter 1956).

3.4. *Finite Life of Excited States*

A thorough analysis of the effect of distant states to which decay is possible would entail a solution of the time-dependent wave equation embracing all the states and radiation fields concerned. This we shall not attempt. We represent the effect of distant states in a phenomenological way by introducing constants γ for exponential decay, $n_t = n_0 e^{-\gamma t}$. The 'lifetime' of the state, $T = 1/\gamma$.

We shall see that T enters into expressions for the form of the resonance curve, the width of which may be used to determine T. The cases of magnetic and electric dipole transitions need to be treated differently, since in the one case all the excited states have the same decay constant, while in the other they do not.

The finite lifetime sets a lower limit to the radio-frequency field strength required to induce a transition (§ 1.2). On the other hand, the stimulation of transitions itself artificially shortens the lifetime of the states, so that the width of the resonance curve is a function of T and of the matrix elements b.

Rather complicated effects are to be expected if T is comparable with the precession time of atomic interactions, the hyperfine structure, for example. It is to be expected in this case that the electron–nuclear coupling would be reduced, so that the peak of a radio-frequency resonance curve for a hyperfine interval

would not exactly correspond with the frequency interval given by the stationary state theory. We understand that the problem has been considered by Serber (1956).

In the following sections the radio frequencies are much larger than the reciprocal lifetimes γ, but a case is cited (§ 3.5) where a perturbing level lies within the distance γ of an excited level.

3.4.1. *Magnetic dipole transitions between excited states.*

Following Brossel and Bitter (1952), we consider the simple case of transitions between a pair of isolated states which have the same lifetime T. We use equation (14) for P_{pq}, the probability of transition in time t. The signal ΔS_{pq} will be some change, usually in the intensity of a beam of light, consequent upon the transition $p \rightarrow q$. Now, of N_p atoms excited and brought under the influence of the radio-frequency field at time t_1, a number $\{N_p \exp[-(t-t_1)/T]\} P_{pq}(t-t_1)$ survive to time $t > t_1$ and, having been transferred to state q, contribute to the signal. The simplicity of this expression rests on the fact that both p and q decay at the same rate. We need to integrate it for all t_1 from $-\infty$ to t. Thus

$$\Delta S_{pq} \propto \int_{-\infty}^{t} \exp\left[\frac{-(t-t_1)}{T}\right] \frac{|b_{pq}|^2}{(\omega-\omega_0)^2+|b_{pq}|^2} \sin^2[\tfrac{1}{2}\{(\omega-\omega_0)^2+|b_{pq}|^2\}^{1/2}(t-t_1)]\, dt_1$$

$$\propto \frac{|b_{pq}|^2}{(\omega-\omega_0)^2+(1/T)^2+|b_{pq}|^2}. \qquad \ldots\ldots(18)$$

In the case of a multiplicity of equally spaced levels, the Majorana formula (17) is used instead of P_{pq}.

Careful experimental studies of line profiles made by Brossel and his colleagues (Brossel and Bitter 1952, Guiochon, Blamont and Brossel 1956) have confirmed the form of equation (18). It has been extensively used in the determination of T through the relation

$$(\tfrac{1}{2}\delta\omega)^2 = |b_{pq}|^2 + (1/T)^2 \qquad \ldots\ldots(19)$$

which gives the line width $\delta\omega$ between points of half-intensity. In so far as $|b_{pq}|^2$ ($\propto H_1^2$) is usually not known absolutely, T is determined by extrapolation to zero radio-frequency field strengths.

3.4.2. *Coherence times.*

T (equation (18)) was interpreted as the lifetime of the excited atoms until the Paris group discovered that its value depended on whether multiple scattering of photons was taking place in the resonance chamber (§ 4.4.2). It is now interpreted as a ' coherence time ' which tends to a constant value τ, the true lifetime, as the atomic density in the vapour is reduced to the point at which multiple scattering ceases.

A theoretical analysis of the relation between T and τ has been outlined (Barrat 1958) and a detailed publication is promised. The problem is stated in terms of the probability amplitudes $a_p(t), a_q(t)$ for the states p and q. There will be definite phase relations between the $a_p(t)$ and $a_q(t)$ for the emitting and absorbing atoms, and there arises the possibility of interference phenomena. An analysis

which describes the emitting or absorbing atom as being definitely in one or other of the states p, q (i.e. in terms of $|a_p|^2, |a_q|^2$), would not disclose the interference effects.

The problem is enlightened by considering the following experiment (Barrat and Brossel 1958). Two separate vessels A and B containing mercury vapour at low pressure (single scattering) are placed in the same steady magnetic field H_0 and the same radio-frequency magnetic field. Vapour in A is excited by resonance radiation from a lamp: the fluorescent radiation from A excites the vapour in B. The fluorescent radiation from B is studied to detect radio-frequency resonances in B. It is found that the details of the resonance curve are quite different from the case of single scattering: in particular, $T_B > \tau$. If now a polarizer be inserted between A and B the characteristics of the resonance curve are those of single scattering.

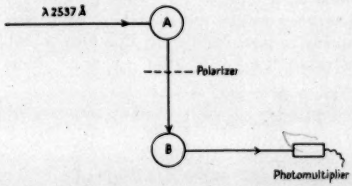


Figure 2. Double scattering of resonance radiation (schematic).

The interpretation of this experiment in the terms proposed above is that, since the atoms are in the same fields, the probability amplitudes $a(t)$ of an exciting atom in A and an excited atom in B develop at the same rate. The phase relations between the $a(t)$ of the atom in A are passed on to the atom in B via the resonance radiation. The coherence time of the excited states is then greater than that of either atom: $T_B > \tau$. With insertion of the polarizer one determines the probability that the photons which pass will have a given polarization, but sacrifices any coherence which existed between different polarizations (Dirac 1947, Ch. I). The phase relations between the emitting atom in A and the excited atom in B become uncorrelated. The measured lifetime will be that of the atom in B.

3.4.3. *Electric dipole transitions between excited states.*

If electric dipole transitions are allowed between a pair of excited states p and q, these will in general decay at different rates. For this reason the treatment of the last section is inapplicable.

We are not aware of any treatment which is valid when the radio-frequency perturbation is not small compared with the natural width of the energy levels. In the case when the perturbation is small, the results of the quantum theory of radiation may be applied (Weisskopf and Wigner 1930, Heitler 1954, p. 186).

Let γ_p, γ_q be the decay constants, and let $\gamma_q > \gamma_p$. Let S_0 be the energy flux per unit area of radiation of angular frequency ω, whose electric vector is parallel to a unit vector $\mathbf{e}$. Then γ_ω, the probability per unit time of a stimulated transition,

may be expressed :

$$\gamma_\omega = \left(\frac{2\pi e^2 S_0}{c\hbar^2}\right)(\gamma_p+\gamma_q)\frac{|r_{pq}|^2}{(\omega-\omega_0)^2+\frac{1}{4}(\gamma_p+\gamma_q)^2} \qquad \ldots\ldots(20)$$

where r_{pq} is the matrix element of $\mathbf{e}.\mathbf{r}$ between states p and q, $\mathbf{r}$ having its usual significance, the sum of the coordinate vectors of the atomic electrons. This expression for γ_ω gives the form of the resonance curve for an isolated pair of excited states, under the approximation that the perturbation is small compared with γ_p or γ_q.

Equation (20) has been derived by Lamb (Lamb and Retherford 1950, Appendix II, Lamb 1952, section L) from the equations of time-dependent perturbation theory (for the particular case that $\gamma_p \simeq 0$) under the approximation that the perturbation is small compared with γ_q.

3.5. *Transitions between Excited States in the vicinity of other States*

Two cases have been investigated in detail. In both cases magnetic dipole transitions were investigated between states p and q, one of which was perturbed by a third state. In the one case the perturbation was by the radio-frequency magnetic field, in the other by a static electric field.

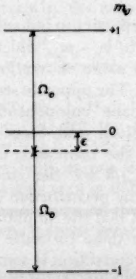


Figure 3. Combined Zeeman and Stark effects in the level $^3P_1$.

The first case (Blamont and Winter 1957, Blamont 1957) concerns the states $m_J = 0, \pm 1$ of the level $^3P_1$ of mercury. The three energy levels are unequally spaced under the influence of static electric and magnetic fields. The Bohr frequencies $(W_q - W_p)/\hbar$ are $\Omega_0 \pm \epsilon$. Blamont and Winter calculate rigorously an expression for the form of the resonance curve, and find excellent agreement with experiment. The positions of the two maxima depend on the radio-frequency field H_1, as one would anticipate from the arguments of § 3.3.2. The perturbation formulae lead to the shifts $\pm(\gamma H_1)^2/4\epsilon$ (matrix element : $\gamma H_1/\sqrt{2}$; $\Delta = 2\epsilon$). Blamont and Winter quote $\pm 4(\gamma H_1)^2/\epsilon$, but this is evidently a misprint, since Blamont's article quotes $\pm(\gamma H_1)^2/4\epsilon$, in agreement with the perturbation formula.

In the second case (Lamb 1952, p. 271, Lamb and Wilcox 1958) magnetic dipole transitions are investigated between the states $m_J = \pm\frac{1}{2}$(α and β) of the level $^2S_{1/2}$ of atomic hydrogen. These are both metastable. The perturbing state is $^2P_{1/2}$, $m_J = \frac{1}{2}$ (e). At about 575 gauss, the states β and e have the same energy. The perturbation between them, V, arises from the electric field which is a consequence of the motion of the atoms through the magnetic field. Under ordinary time-independent theory, the perturbation between β and e would be such that they would never in fact cross: at the field where they would have crossed, each state should be a 50 : 50 mixture of pure S and P state. An entirely different conclusion is reached when one takes account of the different decay constants of β and e.

If $|V| \leqslant \frac{1}{4}\gamma$, where γ is the decay constant of the P state, β and e do in fact cross, and largely retain their S and P characters. Moreover, the point of intersection does not depend on the perturbation V. Hence the peak of the resonance curve is unshifted, and the curve itself is far narrower than would correspond to sharing of the decay constant of state e.

3.6. *Quanta of Different Kinds*

In the foregoing sections we restricted the argument to a uniformly rotating magnetic or electric field. In actual experiments it is usual to employ an oscillating field $H_1 \cos \omega t\, \mathbf{i}$ which corresponds to the superposition of two fields rotating in opposite directions.

The sense of rotation which is effective in inducing the required transition depends on the sign of $W_q - W_p$ relative to $m_q - m_p$, which, in the magnetic case, depends on the sign of the g-factor. The sense of rotation which stimulates absorption of energy also stimulates emission. The opposite sense stimulates neither. We refer to the 'resonant' and 'anti-resonant' components of the oscillating field.

It was first calculated by Bloch and Siegert (1940) that the anti-resonant component causes the peak of the resonance curve to be displaced to higher frequencies from $\omega_0 = (W_q - W_p)/\hbar$ by the amount $|b_{pq}|^2/4\omega_0$, where $\frac{1}{2}b_{pq}$, as before, is the matrix element of the perturbation (see equation (13)). In applying this relation it is to be understood that the amplitude of the oscillating field is twice that of the rotating field. H_1 in these formulae is the amplitude of the rotating component. The Bloch–Siegert correction is sometimes significant in experiments on excited states.

Ramsey (1955) has shown that if rotating fields of several frequencies $\omega_1, \ldots, \omega_i \ldots$ are present simultaneously, the peak of the resonance curve at ω_1 near ω_0 is shifted as before by the amount $\omega_1 - \omega_0 = |b_i|^2/2(\omega_0 - \omega_i)$ for each ω_i, provided that $|b_1| \ldots |b_i| \ldots \ll \omega_0 - \omega_i$. This expression includes the Bloch–Siegert case $\omega_i = -\omega_0$.

A type of multiple quantum transition, different from that discussed in § 3.3.2, has been investigated by Brossel, Margerie and Winter (1955). Transitions between a given pair of states are induced by the simultaneous action of radio-frequency fields of different frequencies ω and ω'. The various possibilities are governed by the equation

$$f\omega + g\omega' = \omega_0,$$

where f and g are positive or negative integers.

Conservation of angular momentum is satisfied by the selection of photons of appropriate polarization. The radio-frequency field must have a component along the axis of quantization as well as perpendicular to it, so that there exist σ^+, σ^- and π photons having 1, -1, and 0 units of angular momentum respectively, in the direction of the axis. An atomic transition $\Delta m = +1$, for example, may then be stimulated by the absorption of one photon, σ^+; or two, σ^+ and π; or three, σ^+, σ^- and π; etc. (figure 4).

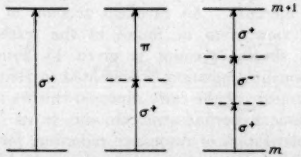


Figure 4. Simultaneous action of quanta of different kinds.

One finds for these resonances the same features as in the case of multiple quantum transitions treated earlier (§ 3.3.2), namely, Lorentz line shape, reduction in line width according to the number of quanta participating, intensity proportional to $|b|^{2f}|b'|^{2g}$ and shift of the peak of the resonance curve. These shifts can be quite large. They are given by the expression

$$\Omega - \omega_0 = \frac{|b|^2}{2(\Omega-\omega)} + \frac{|b|^2}{2(\Omega+\omega)} + \frac{|b'|^2}{2(\Omega-\omega')} + \frac{|b'|^2}{2(\Omega+\omega')} \qquad \ldots\ldots(21)$$

where $\Omega = f\omega + g\omega'$ at resonance, $\omega_0 = (W_q - W_p)/\hbar$ and $\frac{1}{2}b, \frac{1}{2}b'$ are the matrix elements as defined in equation (13) for the two radio-frequency fields which together induce the transition $\Delta m = +1$. The formula is deduced for linear radio-frequency fields. It is valid so long as one of the quantities $|b|^2/(\Omega-\omega)\ldots$ is small compared with Ω. Terms with zero in the denominator are to be discounted.

Although there may be components of the radio-frequency field along the axis, the matrix elements into which these enter are zero, and they do not appear in equation (21).

In an experiment on sodium in the ground level, Brossel, Margerie and Winter (1955) observed shifts of resonance peaks which agreed well with the predictions of (21).

§ 4. Double Resonance Experiments

4.1. *Resonance Radiation*

The phenomena of optical resonance radiation attracted a great deal of attention in the nineteen-twenties and thirties. Wood's early experiments with sodium vapour were followed by quantitative studies of the dependence of the intensity and polarization of the fluorescent radiation on such variables as the vapour pressure of the fluorescent substance, the pressure of foreign gases, the presence of magnetic fields, the geometry of the absorption bulb, and the spectral distribution of the exciting radiation.

It will be recalled that essential conditions for the observation of resonance radiation are low vapour pressure of the substance concerned (<1 micron of mercury) and illumination by light whose frequency corresponds exactly with an allowed transition in the substance, normally from the ground level. This light may, but need not, originate from a lamp containing the particular substance. Self-reversal in the light source must be avoided. Low pressure of foreign gas in the resonance vessel is usually desirable, but is not always essential. Resonance radiation has been observed in many substances, although most work has been done on sodium and mercury. An excellent account of the subject from the experimental point of view is to be found in the textbook by Mitchell and Zemansky (1934). A shorter account is given by Pringsheim (1949). The quantum theory of resonance radiation is presented in Heitler's book (1954).

In studying the literature of the early nineteen-thirties it is to be remembered that the theory of atomic hyperfine structure was in its infancy. Some of the experiments on the polarization of resonance radiation, for example, which were difficult to interpret at the time, are immediately explicable in terms of nuclear spin.

Of particular importance in connection with double resonance work is an understanding of the distribution in space of resonance radiation, and its polarization. While quantum mechanical calculations are necessary for detailed studies, a model based on classical physics can be very helpful. This we next consider. With its aid we realize, for example, the important practical point that the angular size of light cones can be quite large before the efficiency of double resonance experiments is seriously impaired.

4.2. *Spatial Distribution of Resonance Radiation : Classical Model*

We first represent the atom by an isotropic electric oscillator. Let this be excited by a polarized electromagnetic wave travelling in a particular direction. The oscillator will be stimulated to radiate with the anisotropic distribution characteristic of a dipole.

An external magnetic field will, in general, exert a couple on the oscillating dipole, and the resulting precessional motion will cause a re-distribution of the scattered radiation. But in the important special case of a field parallel to the axis of the dipole there will be no couple, no precession and no change in the scattered radiation.

We wish now to examine the effect of an additional magnetic field, rotating in a plane perpendicular to the static field. We consider the special case cited above. Resolve the oscillating electric dipole into two opposed circular currents, that is, two oppositely directed magnetic dipoles. We endow these with angular momenta parallel to the dipole moments, so that the two angular momenta are also opposed. The magnetic dipoles now precess in the same sense about the static field. If the additional magnetic field rotates at the precessional frequency and in the correct sense, there will occur energy exchanges between the dipoles and the rotating field : that is to say, the dipoles may change their direction with respect to the static field. In this way the rotating magnetic field can bring about a re-distribution of the radiation from the oscillator, and resonance between the rotating field and the precession of the atom may thereby be detected.

In an actual atom the spin–orbit or hyperfine structure interactions may exert couples on the oscillator, which is then no longer isotropic. The distribution of radiation is different in this case, but the possibility still exists of changing the distribution by means of a rotating magnetic field of the correct frequency. This may be the precessional frequency in an external field, as before, or it may be any of the precessional frequencies associated with the spin–orbit or hyperfine structure interactions in zero field.

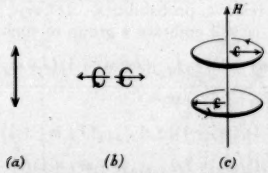


Figure 5. Resolution of an oscillating electric dipole: (*a*) oscillating electric dipole; (*b*) equivalent pair of spinning magnetic dipoles; (*c*) precession of the magnetic dipoles in a static field H.

4.3. *Intensity and Polarization of Resonance Radiation: Quantum Theory*

For many purposes it is satisfactory to treat the process of resonance fluorescence as the succession of two independent processes, excitation and decay. This treatment we shall shortly illustrate. It is inadequate when the natural width of the levels is comparable with the distance between them, for then the possibility arises of interference between excitations to different states. The case has been discussed by Hawkins (1955) and by Serber (1956). When the levels are well separated the interference effects are negligible, and the simple treatment suffices.

Knowledge of the probability for excitation to a particular level demands in principle a calculation of the matrix elements of the electric dipole moment. Absolute calculations are generally quite impracticable: we use instead the measured f-value of the transition, which gives the total cross section for excitation, and the angular factors in the matrix elements, which give the distribution between different states.

The f-value (oscillator strength) of an atomic transition is a measure of the ability of an atom to absorb radiant energy relative to the ability of a classical oscillating electron. It is in fact connected with the probability for spontaneous decay of the excited level A_{kj} by the relation

$$f_{jk} = \frac{mc\lambda_{kj}^2}{8\pi^2 e^2}\frac{g_k}{g_j}A_{kj} = 1{\cdot}51\frac{g_k}{g_j}\lambda_{kj}^2 A_{kj} \quad (\lambda_{kj} \text{ in cm})$$

where g_k, g_j are the statistical weights of the levels k, j.

The integrated cross section for absorption of radiation by a classical oscillating electron is a universal constant, $\pi e^2/mc$. We therefore have, for absorption by a transition of oscillator strength f,

$$\int \sigma_\nu \, d\nu = \pi e^2 f/mc \qquad \ldots\ldots(22)$$

where σ_ν is the cross section at frequency ν. If the atoms are illuminated by $n_\nu d\nu$ photons/cm² sec in the spectral range $d\nu$, the number of photons/sec absorbed per atom will be

$$\int n_\nu \sigma_\nu d\nu = n_\nu \pi e^2 f/mc = 2{\cdot}64 \times 10^{-2} n_\nu f \qquad \ldots\ldots(23)$$

if n_ν is a constant over the absorption region. Values of f are to be found in Landolt–Börnstein's tables (1950).

We now quote the relative probabilities, $A(J, m_J\,;\, J', m_J')$, for transitions between different states. f will embrace a group of such transitions.

$$A[J, m_J\,;\, J, (m_J \pm 1)] : A_{J,J}(J \pm m_J + 1)(J \mp m_J)$$

$$A[J, m_J\,;\, J, m_J] : 2A_{J,J} m_J^2$$

$$A[J, m_J\,;\, (J+1), (m_J \pm 1)] : A_{J,J+1}(J \pm m_J + 1)(J \pm m_J + 2)$$

$$A[J, m_J\,;\, (J+1), m_J] : 2A_{J,J+1}(J + m_J + 1)(J - m_J + 1). \qquad \ldots\ldots(24a)$$

In using the formulae for calculating the intensity of light in absorption or emission, it is to be noticed that the spatial distribution of π light ($\Delta m = 0$) is different from that of σ light ($\Delta m = \pm 1$). For this reason, when the intensities of π and σ light are being compared for a direction at right angles to the axis of quantization, the transition probabilities for σ light (given below) are to be multiplied by $\frac{1}{2}$. It will be noticed that the totals of π light and of σ light in this direction are always equal.

For the Zeeman components in hyperfine structure we use the above formulae with J, m_J substituted by F, m_F. The coefficients $A_{F,F}$, $A_{F,F+1}$ are then related as follows:

for $J \to J$,

$$A_{F,F} = A_{J,J}\left\{\frac{J(J+1)+F(F+1)-I(I+1)^2}{F(F+1)}(2F+1)\right\}$$

$$A_{F,F+1} = A_{J,J}\left\{\frac{(J+F+I+2)(J-F-I-1)(J-F+I)(J+F-I+1)}{(F+1)}\right\};$$

for $J \to J+1$,

$$A_{F,F} = A_{J,J+1}\left\{\frac{(J+F+I+2)(J+F-I+1)(J-F+I+1)(J-F-I)}{F(F+1)}(2F+1)\right\}$$

$$A_{F,F+1} = A_{J,J+1}\left\{\frac{(J+F+I+3)(J+F+I+2)(J+F-I+2)(J+F-I+1)}{(F+1)}\right\}. \qquad \ldots\ldots(24b)$$

These formulae have been gathered from the book by Condon and Shortley (1951, pp. 63, 238, 387, 424).

The formulae cannot be used for hyperfine structure if the atom is in a magnetic field which is strong for the excited, and weak for the ground level. In this case each ground state (F, m_F) is treated as a sum of states $(J, m_J\,;\, I, m_I)$ with probability amplitudes which are conveniently given by Condon and Shortley, p. 73.

By way of example, we now apply the formulae to the resonance line of mercury, $6^1S_0 - 6^3P_1$, λ2537Å. The cases $I = 0$ (even isotopes) and $I = \frac{1}{2}$(^{199}Hg) are illustrated in figure 6, where the numbers which break the transition lines are ratios of transition probabilities, $A(J, m_J\,;\, J', m_J')$ or $A(F, m_F\,;\, F', m_F')$.

Normally, all the ground states of a given isotope will be equally populated (apart from the possibility of a very small Boltzmann factor, which we ignore).

We examine first the effect of illumination whose frequency distribution is uniform over the whole of the absorption region.

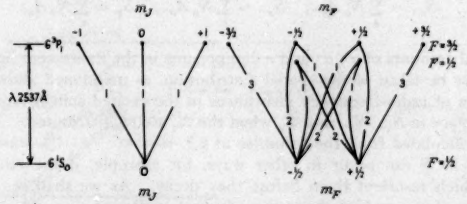


Figure 6. Transition probabilities in the resonance line of mercury: (*b*) even isotopes: $I = 0$. (*c*) ^{199}Hg: $I = \frac{1}{2}$.
Relative intensities for observation perpendicular to the axis of quantization are obtained by introducing the factor $\frac{1}{2}$ for the transitions $\Delta m = \pm 1$.

The relative populations of the excited states will then be the sums, for each state, of the excitation probabilities for the given type of exciting radiation (bearing in mind the factor $\frac{1}{2}$ for σ radiation where appropriate). Thus, for excitation in a direction perpendicular to the axis of quantization, we find the following distribution of population.

	even isotopes			^{199}Hg					
	$J = 1$			$F = \frac{3}{2}$				$F = \frac{1}{2}$	
	-1	0	$+1$	$-\frac{3}{2}$	$-\frac{1}{2}$	$+\frac{1}{2}$	$+\frac{3}{2}$	$-\frac{1}{2}$	$+\frac{1}{2}\,m$
Excitation by									
π light	0	1	0	0	2	2	0	1	1
σ^- light	$\frac{1}{2}$	0	0	$\frac{3}{2}$	$\frac{1}{2}$	0	0	1	0
σ^+ light	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{3}{2}$	0	1
natural light	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	2	2

The examples bring out the following facts, which are true generally:

(i) When the values of J (or F) differ between the ground and excited levels, population differences may be induced by light of any polarization, or by natural light. Differences are most pronounced when π light is used.

(ii) When the values of J (or F) are the same in the excited as in the ground level, population differences are induced only by σ^- or σ^+ illumination.

It may be verified that isotropic excitation by natural light results in equal population of all the excited states.

We now consider the case when the exciting radiation is non-uniform in frequency distribution, for example, when the hyperfine structure is partially resolved in the light from a source. The transition probabilities are now to be multiplied by the relative intensities of the light in the difference regions of absorption. The result is that the excited states can be unequally populated, even when the illumination is isotropic and unpolarized. We consider a particular example in § 4.6.

The populations N_e of the excited states having been found, the quantities

$$S_{\sigma^+} = \sum_e N_e A_{\sigma^+}, \quad S_{\sigma^-} = \sum_e N_e A_{\sigma^-}, \quad S_\pi = \sum_e N_e A_\pi$$

give the total amounts of σ^+, σ^- and π components in the fluorescent light. Proper account must be taken of the spatial distribution, as mentioned above.

Detection of radio-frequency resonances in the excited states rests on changes which take place in $S_{\sigma^+}, S_{\sigma^-}$ and S_π when the N_e are re-distributed. The changes ΔS can be calculated from the formulae of § 3.

Changes in S can occur in other ways, for example, if the atoms undergo collisions which re-orient them before they decay. As we shall see later, states with $J = 0$ are less sensitive than others to such collisions.

The smallest signal which can be detected depends on the sources of noise in the experiment. Frequently, the most important source is the shot noise in the emission of electrons from the cathode of a photomultiplier. In this case we can calculate the signal to noise ratio as follows :

Let i_s be the primary current at the cathode of the photomultiplier due to the fluorescent light S. Stray light may contribute a further current i_0. The shot noise is constituted by the random fluctuations in the n electrons which, in observation time τ, constitute the current $i = i_s + i_0$. The r.m.s. noise current i_n, averaged for time τ, is the current associated with $n^{1/2}$ electrons, i.e.

$$i_n = n^{1/2} e/\tau = (ie/\tau)^{1/2}, \quad \text{since } i = ne/\tau.$$

The signal current $i_{\Delta s}$, is $(\Delta S/S) i_s = \beta i_s$. Thus

$$\frac{i_{\Delta s}}{i_n} = \frac{\beta i_s}{(ie/\tau)^{1/2}}. \qquad \ldots\ldots(25)$$

The fraction β depends only on the relative populations N_e, the radio-frequency transition probabilities, and the relative probabilities for decay. It does not depend on the spectral density of the exciting radiation, provided this is uniform over the region of absorption. Nor does β depend on the f-value, or on the density of absorbing atoms N_a.

i_s, on the other hand, depends on n_ν, f and N_a, also on the fraction of the fluorescent light which is received by the photomultiplier ϕ and on the efficiency of the photocathode η. For if, as in equation (23), $n_\nu\, d\nu$ is the number of photons per cm^3 in the spectral range $d\nu$, then the number of atoms excited per second is $(\pi e^2/mc)\, n_\nu f N_a l$, provided that the vapour is optically thin, where l is the length of vapour column of N_a atoms per cm^3. We then have $i_s = (\pi e^2/mc)\, n_\nu f N_a l \phi\eta$.

4.4. *Double Resonance Experiments in Magnetic Fields*

We now give brief accounts of experiments which have been performed.

4.4.1. *The experiment of Brossel and Bitter.*

The ' double resonance ' method was introduced by Brossel and Bitter (1952), following the prediction of Brossel and Kastler (1949). The excited level 6^3P_1 of mercury was studied in a magnetic field. π excitation with the resonance line $\lambda 2537$Å was chosen, i.e. the light was incident on the resonance cell at right angles to the magnetic field, and was polarized with its electric vector parallel to the field.

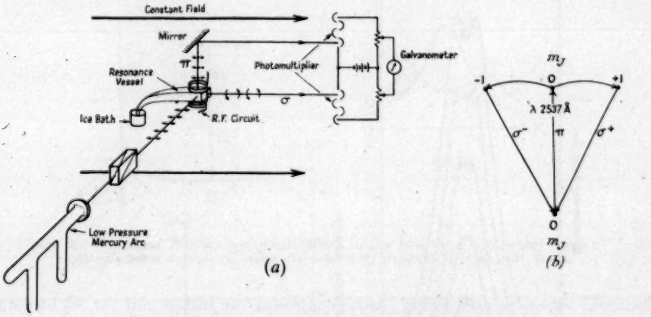


Figure 7. The experiment of Brossel and Bitter.

Under this excitation the state $m_J = 0$ alone is excited. The radio-frequency magnetic field at fixed frequency ν was applied by a coil surrounding the resonance vessel, and the ' static ' magnetic field varied through the resonant value H_0 given by $h\nu = g_J \beta H_0$. It is clear from figure 7(*b*) that radio-frequency resonance will be accompanied by an increase in σ, and a decrease in π light. These changes were detected by photomultipliers which received light in the direction of, and perpendicular to, the field, respectively. The currents were set in opposition in order to compensate fluctuations in the light source, and at the same time to double the signal. The difference current was observed on a galvanometer.

Brossel and Bitter studied in detail the line shape of the radio-frequency resonances, and in particular its dependence on the radio-frequency field strength. As is shown in § 3.4.1 it is possible by extrapolation to obtain a line width, and hence a time, which was interpreted as the lifetime of the atoms. Further studies, described in the next section, have revealed that the ' coherence time ' obtained in this way is not identical with the lifetime but is asymptotic to it at low vapour pressures.

Figure 8(*a*) shows a set of resonance curves obtained with various amplitudes of radio-frequency field. The double peaked curves obtained at the higher field strengths are predicted by the Majorana formula (equation (17)), which is appropriate to this case.

Figure 8(*b*) demonstrates displacement of the resonances by the Bloch–Siegert effect (§ 3.6).

From their measurements Brossel and Bitter deduced the value

$$g_J = 1{\cdot}4838 \pm 0{\cdot}0004$$

for the level 6^3P_1 in mercury. The significance of the experiment lies not so much in this determination as in the demonstration of the feasibility of the method. From this work has stemmed the investigation not only of other excited states, but also of the widening and important field of ' optical pumping '.

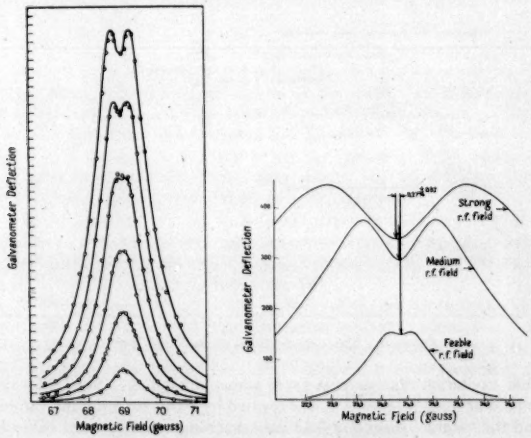


Figure 8. Resonance curves obtained by Brossel and Bitter: (*a*) resonances at different r.f. field strengths; (*b*) the Bloch–Siegert effect.

4.4.2. *Further studies of excited states in mercury.*

Blamont and Brossel (1956) studied the Stark effect of the level 6^3P_1 in mercury. The experimental arrangement resembled that of Brossel and Bitter, with the addition of a pair of plane electrodes in the vapour for the establishment of static electric fields of the order of 50 kv cm^{-1}. The difficulty of finding electrodes which could be used in a region of strong radio-frequency fields was overcome by using a semiconductor coated on to Pyrex (Blamont 1957).

In contrast with the Zeeman effect, the Stark effect is independent of the sign of m_J (or m_F). The intervals between $m_J = 0$ and ± 1, therefore, which are equal under a magnetic field alone, become unequal when an electric field is simultaneously applied (figure 9(*a*)): the single peak ν_0 (figure 9(*b*)), which denotes two coincident radio-frequency resonances, separates into its two components ν_1 and ν_2. The interpretation of these curves is not quite straightforward, since

doubling of the peak can occur even when there is no electric field if the radio-frequency field is strong enough (see figure 8). Blamont's detailed theory of the line shape (1957, see also § 3.5) is well supported by the measurements.

Guiochon, Blamont and Brossel (1956, 1957) also Boutron, Barrat and Brossel (1957) observed and studied an unexpected narrowing of the radio-frequency resonances with increase in vapour pressure of the mercury. One would expect that the shorter mean free path associated with higher vapour pressures would shorten the lifetimes of the excited atoms and lead to wider radio-frequency resonances, but this is contrary to their observations. They describe the times

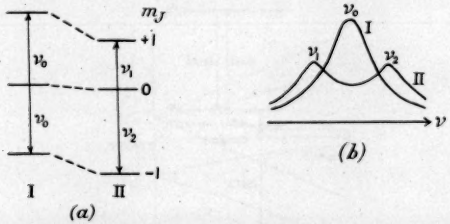


Figure 9. Combined Zeeman and Stark effects in the level $^3P_1$: (*a*) energy levels; (*b*) resonance curves, (I) magnetic field, (II) magnetic and electric fields.

obtained for the line widths as ' coherence times ' rather than as atomic lifetimes, and relate the longer ' coherence times ' to multiple scattering of photons which takes place when the vapour pressure is increased. The phenomenon is discussed in § 3.4.2. Guiochon *et al.* support their interpretation of line narrowing by numerous measurements of coherence times as a function of vapour pressure in specimens of mercury of different isotopic composition : an increase in the number of atoms of a different sort will not lead to increased scattering. Measurements of ' lifetimes ' by the classical methods of depolarization of resonance radiation in magnetic fields lead to ' coherence times ' in agreement with the values obtained from double resonance experiments (Barrat 1957). The true lifetime of the level 6^3P_1 of mercury they find to be $1{\cdot}18 \times 10^{-7}$ sec.

The level 7^3S_1 of mercury has been studied by Brossel and Julienne (1956). It is possible to excite this level stepwise by the illumination of mercury vapour with radiation 2537Å and 4358Å simultaneously, but the yield of doubly excited atoms is rather poor. Far greater yields are obtained by making use of the metastability of the level 6^3P_0, to which atoms in the 6^3P_1 level are transferred by collision with molecular nitrogen. Excitation to 7^3S_1 is achieved by illumination with radiation 4046Å rather than 4358Å.

Brossel and Julienne therefore added nitrogen at a pressure of several millimetres of mercury to the absorption vessel containing mercury vapour. The exciting radiations, wavelengths 2537Å and 4046Å from a mercury lamp, were unpolarized, but directed along the applied magnetic field, corresponding therefore to σ excitation. Radio-frequency resonance was detected by changes in the

intensity of the σ and π components of the emitted radiation 4358Å. Resonance peaks at about 53 gauss were observed at a frequency of 148 Mc/s.

Two points are of particular interest in this work: One is the successful application of a radio-frequency magnetic field of high intensity and high frequency (about 4 gauss at 150 Mc/s) to a gas at several millimetres pressure without the stimulation of an ordinary discharge. This was accomplished by arranging that the radio-frequency magnetic field in the resonance vessel was uniform, and almost entirely free of the electric field. These conditions were established at the short-circuited end of a tuned Lecher wire system formed from flat metal strip. The

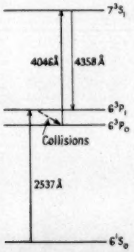


Figure 10. Stepwise excitation from the metastable level in mercury.

second point of interest is that the 7^3S_1 states were able to survive in the presence of the nitrogen. The line width of the resonance curves published by Brossel and Julienne is about 20 gauss, corresponding to a time of about 3×10^{-9} sec, which gives the order of magnitude of the lifetime. Attempts are now being made to investigate how the pressure of nitrogen affects the line width.

Kastler, Brossel and their pupils have applied double resonance techniques in the process known as ' optical pumping ' to sodium in particular, and have studied in great detail the phenomena of multiple quantum transitions discussed in § 3.3.2 of this report. Since these experiments are concerned with the ground states they will not be described here. They form part of the subject matter of Kastler's Holweck Lecture (1954) and of a number of papers given at the Magnetic Resonance Conference in Paris in 1958 (Kastler 1958) where further references may be found.

Bogle, Dodd and Maclean (1957) have studied the level 6^3P_1 of mercury with a view to determining the hyperfine structure coupling constants of the odd isotopes. The levels $F = \frac{3}{2}$, $F = \frac{1}{2}$ are separated in zero field by about 20 000 Mc/s. In a magnetic field of 2–3 kilogauss the Zeeman components are separated by intervals of about 4000 Mc/s, but these are no longer equal (figure 11), since the electron–nuclear coupling is beginning to break down. From the differences, the hyperfine structure coupling constant can be deduced.

The method of illumination and detection resembled that of Brossel's work, but whereas Brossel and his colleagues used frequencies in the range of conventional high-frequency oscillators, Bogle *et al.* worked at 4250 Mc/s with klystron

and waveguide techniques. A resonance cavity of size appropriate to the wavelength, about 7 cm, allowed the insertion of a fused quartz resonance cell of reasonable dimensions.

Of the two microwave resonances observed, one was interpreted as the single resonance expected from the even isotopes (this occurred at the g_J-value measured by Brossel and Bitter) and the other as one of the four resonances expected from the isotope 199. The hyperfine structure A-factor was determined with a precision to about 2%, and agrees with optical determinations of higher precision. As the authors point out, greater accuracy could be obtained by working at higher fields where the Back–Goudsmit effect is more pronounced.

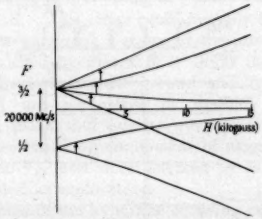


Figure 11. Zeeman effect in the hyperfine structure of the level $^3P_1$ of ^{199}Hg ($I = \frac{1}{2}$).

4.4.3. *Excited states in potassium and sodium: non-zero fields.*

Ritter and Series (1957) observed resonances in the second excited P levels of potassium, $5P_{3/2}$ and $5P_{1/2}$. The experimental features are more conveniently deferred to the next section on resonances in zero field. Here we wish to describe the work in non-zero fields.

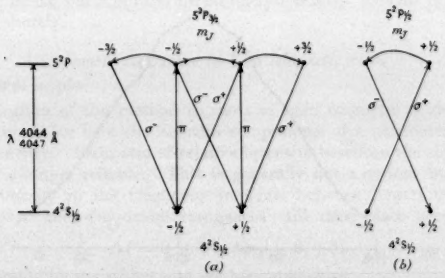


Figure 12. Double resonance in the term 5^2P of potassium: (*a*) $5^2P_{3/2}$: excitation by π, detection by σ radiation; (*b*) $5^2P_{1/2}$: excitation by σ^+, detection by σ^- radiation. For (*c*), (*d*), (*e*) see overleaf.

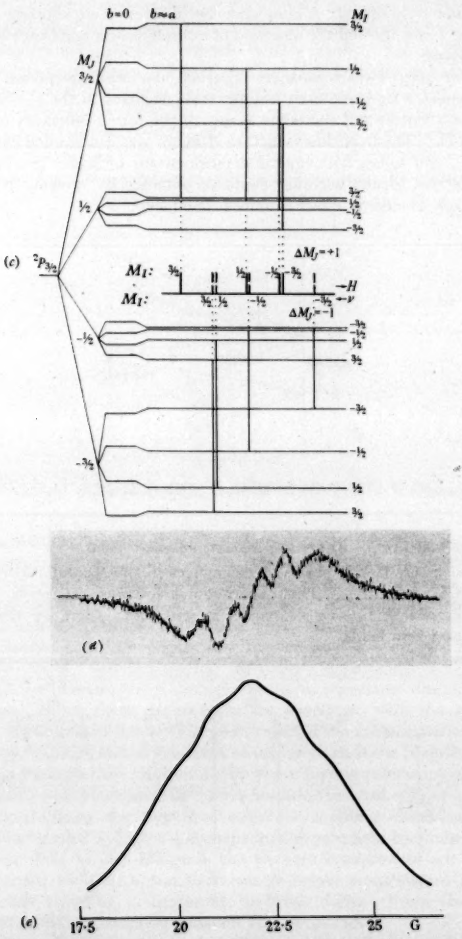


Figure 12 (*contd.*). Double resonance in the term 5^2P of potassium : (*c*) hyperfine structure of $^2P_{3/2}$ in a ' strong ' magnetic field ; (*d*) experimental curve showing hyperfine structure in $5^2P_{3/2}$; the phase-sensitive method of detection displays a differentiated resonance curve ; (*e*) the experimental curve integrated.

The magnitude of the hyperfine structure in these levels is exceptionally small, so that in magnetic fields of 25 gauss the electron–nuclear interaction in the $P_{3/2}$ level is almost entirely broken down. At these fields, under excitation by π light, magnetic resonance curves were obtained in σ^+ and σ^- light, combined and separately. The curves show hyperfine structure from which the interaction constants and nuclear moments were deduced.

These resonances in the $P_{3/2}$ level were observed also at weaker magnetic fields where the I–J interaction was only partially broken down. It was possible to plot the position of two components over the whole range of magnetic fields from strong to zero. Comparison of the results with the stationary-state theory of the excited level (§ 2.1) showed good, but not perfect agreement. Particularly noticeable was that the degree of decoupling of I and J was affected by the strength of the radio-frequency field. This effect could possibly find an explanation in the terms proposed in the last paragraph but one of the introductory § 3.4.

The $P_{1/2}$ level was studied by means of σ^+ light in excitation, σ^- in decay. (When J does not change between the ground and excited levels, π light and undifferentiated σ light are ineffective in double resonance experiments.) Discrimination by Nicol prism and quarter-wave plate between the two senses of rotation requires a non-vanishing component of magnetic field along the light paths. Resonances in the $P_{1/2}$ level were detected by Ritter and Series in experiments which were regarded as exploratory.

The experimental values of the hyperfine coupling constants refer to the more abundant isotope, ^{39}K. The values of $a_{3/2}$ and $b_{3/2}$ agreed with those derived from zero field measurements. Values of the nuclear magnetic moment calculated from $a_{3/2}$ and $a_{1/2}$ agreed with more precise nuclear resonance measurements. The nuclear electric quadrupole moment was calculated from $b_{3/2}$ (§ 2.2).

Dodd (private communication) has measured hyperfine structure in the level $3^2P_{3/2}$ in sodium in magnetic fields about 2300 gauss, at a frequency about 4250 Mc/s. The apparatus was similar to that used in the study of the level 6^3P_1 in mercury (§ 4.4.2). The values of the coupling constants are substantially in agreement with those obtained by Sagalyn (1954) and by Perl, Rabi and Senitzky (1955) in zero magnetic fields, but b is determined unequivocally. Details of the work will be published shortly.

4.5. *Double Resonance in Zero Magnetic Field*

4.5.1. *General principles.*

The application of the method to cases of zero magnetic field needs a little further discussion, for here the Zeeman components of a particular level are not separated in energy. Estimates of relative intensity based on the simple approach of § 4.3 are no longer reliable. This is generally not a serious matter if one is interested primarily in the frequency intervals between hyperfine components. Magnetic dipole radio-frequency resonances still take place according to the selection rules

$$\Delta F = 0, \pm 1\,; \quad \Delta m_F = 0, \pm 1,$$

and are detectable by the techniques we have described.

Since there will always be other levels in the vicinity of the particular pair one is investigating, estimates should be made according to the principles of § 3.3.2

of shifts in the resonance frequency. These will usually be a small fraction of the line width provided the latter is small compared with the resonant frequency. When this condition is not satisfied it will frequently be found that the approximations required by the perturbation treatment are not valid. More theoretical work is required before the fullest information can be obtained from these double resonance measurements. In the meantime, the deduced interaction constants are usually quoted with rather less precision than would be possible if one had a better understanding of the resonance process.

The effect of small stray fields needs to be considered in zero-field experiments. This we now examine.

In a field which is strictly zero the axis of quantization for energy eigenstates is determined with reference to the polarization of the light beam : plane polarized light is termed π if the axis is parallel to the electric vector, σ if perpendicular ; for circularly polarized light the axis is in the direction of propagation of the light. This terminology is consistent with the selection rules $\Delta m = 0, \pm 1$ for π and σ radiation, respectively. Let the states be labelled (F, m_F) with respect to the axis z so determined.

If a field is now established in a direction z' different from z, the states (F, m_F) are no longer energy eigenstates. The useful representation for the purpose of measuring energy intervals is (F', m'_F) with reference to z', and in this representation each state is a mixture of the states (F, m_F). The population distribution in (F', m_F') is quite different from that in (F, m_F). Consequently a double resonance experiment designed on the basis of the (F, m_F) representation may fail entirely.

The transition from the one representation to the other is not quite sudden if the states have a finite lifetime. For if the atoms decay before they have executed a significant precession in the z' field, the direction z' is to that extent imprecisely defined for the atom. We deduce, therefore, that a stray field H' can be tolerated provided that

$$g_F \beta H' < 1/\tau$$

where τ is the lifetime of the state, i.e. the separation of the Zeeman components must not exceed the natural width of the energy levels.

For ordinary excited levels, the requirement of a sufficiently small field is not severe : a lifetime of 5×10^{-8} sec, for example, corresponds to a natural width of about 3 Mc/s, and (for $g_F = 1$) to a magnetic field of about 2 gauss. In such a case the earth's field need not be compensated. For metastable states, on the other hand, the earth's field may be uncomfortably great. In experiments on the level 4^3P_1 in zinc, for example ($\tau \simeq 3 \times 10^{-5}$ sec), Böckman, Krüger and Recknagel (1957) report that they were obliged to reduce stray fields to about 0·05 gauss.

4.5.2. *Experiments.*

A number of authors have performed these experiments with the principal object of measuring hyperfine structures and determining nuclear electric quadrupole moments in elements where they were not known. The quadrupole interaction is only measurable in levels for which $J > \frac{1}{2}$: the $P_{3/2}$ levels of the alkalis, therefore, lend themselves to such determinations. In the first excited $P_{3/2}$ levels the ratio of natural width to hyperfine structure is less favourable than in the second, consequently more work has been done on the second excited levels. For these

levels, however, the oscillator strengths are smaller than in the first by a factor of about 100. This affects the strength of the exciting radiation in the light source and the cross section for excitation, with consequent reduction in the intensity of the fluorescent light.

The following levels have been investigated:

sodium-23 (ground level $3^2S_{1/2}$):	$3^2P_{3/2}$:	Sagalyn (1954)
	$4^2P_{3/2}$:	Krüger (1958)
potassium-39 (ground level $4^2S_{1/2}$):	$5^2P_{3/2}$:	Ritter and Series (1957)
rubidium-85 and 87 (ground level $5^2S_{1/2}$):	$6^2P_{3/2}$:	Meyer-Berkhout (1955)
caesium-133 (ground level $6^2S_{1/2}$):	$7^2P_{3/2}$:	Althoff (1955)
	$7^2P_{1/2}$:	Bucka (1958) (see § 4.6)
zinc-67 (ground level 4^1S_0):	4^3P_1:	Böckman, Krüger and Recknagel (1957)

The experiments are not sufficiently different to justify separate discussion. Light from the source (figure 13) is polarized and focused on to the resonance

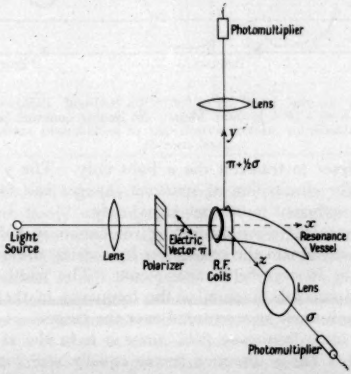


Figure 13. Double resonance in zero magnetic field.

vessel, which itself is focused on to a photomultiplier. In most cases the resonance vessel is horn-shaped for the avoidance of back-reflected light from its walls. The glass, or other material, need not be of particularly high optical quality since good image formation is not necessary. Radio-frequency currents of several amperes are usually required in the coils, and this frequently causes an objectionable gas discharge in the resonance vessel. Partly for this reason most workers have found it necessary to free their samples from foreign gas, either by getters, or by careful re-distillation before sealing in, or by the continuous pumping of a crude atomic beam. When sealed-in samples have been used, the temperature must be controlled in order to establish the necessary vapour pressure. Streams of hot air are often adequate.

Foreign gas is harmful also in its disorientation of excited states through collisions. Second excited states seem particularly sensitive to foreign gas; on the other hand polarized resonance radiation has been observed from the first excited state in sodium in the presence of foreign gases at pressures of several tenths of a millimetre of mercury (Brossel, Margerie and Kastler 1955). S states are especially stable against disorienting collisions.

With the geometrical arrangement of figure 13 the z direction is the axis of quantization. A photomultiplier in this direction receives σ light only. A photomultiplier in the y direction receives a mixture of π and σ light, and might be

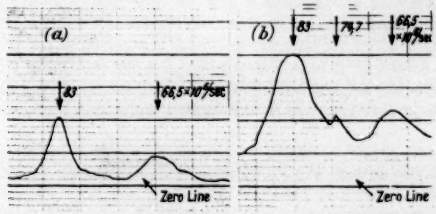


Figure 14. Resonances in the level $7^2P_{3/2}$ of ^{133}Cs (Althoff 1955): (*a*) $(F=5)\rightarrow(F=4)$ (83 Mc/s) and $(F=4)\rightarrow(F=3)$ (66·5 Mc/s); (*b*) double quantum jump $(F=5)\rightarrow(F=3)$ $(2\times74{\cdot}7$ Mc/s).

preceded by an analyser to transmit the π light only. The y and z signals are sometimes opposed for elimination of spurious changes and for doubling of the effect, but it is often sufficient to use one signal only.

Whereas it is convenient when examining resonances in non-zero fields to fix the radio frequency and obtain the resonances by varying the field, it is necessary to vary the frequency in zero-field experiments. The width of the resonance curves is often an appreciable fraction of the frequency of the peak: the radio-frequency field strength must be monitored over the range.

In figure 13 the radio-frequency field, since it is in the x direction, induces transitions $\Delta m=\pm1$. The y direction serves equally well for $\Delta m=\pm1$. The transitions $\Delta m=0$ are induced when the radio-frequency field is in the z direction.

The transitions $\Delta m_F=0,\pm1$ between a given pair of hyperfine levels $F, F+1$ coincide when the magnetic field is zero or very weak. The radio-frequency spectrum normally corresponds to the intervals between the hyperfine levels $F, F+1, F+2\ldots$. Double quantum jumps $\Delta F=\pm2$ are not infrequent (figure 14): indeed, they are to be expected if the resonance line width is comparable with the frequency intervals, for then the radio-frequency perturbation is sufficiently strong to favour the double quantum transition (§ 3.3.2).

4.6. *Excitation by Unpolarized Light in Zero Magnetic Field*

In the following method, due to Bucka (1955, 1956, 1958), double resonance experiments may be performed in zero magnetic field with unpolarized light incident isotropically, or from some limited cone.

A simple example will illustrate the principle. Consider the transition between the ground level $J = 0$ and an excited level $J = 1$ of a substance for which $I = \frac{1}{2}$ (figure 15). Let the two hyperfine components be resolved in a light source : their intensities will be in the ratio 2 : 1.

Now if atoms in the ground level were excited by isotropic radiation of uniform spectral distribution, all the excited states would be equally occupied. Irradiation from the given light source, on the other hand, will over-populate by the factor 2 all the states which belong to $F = \frac{3}{2}$ compared with those of $F = \frac{1}{2}$. Radio-frequency resonances between $F = \frac{3}{2}$ and $F = \frac{1}{2}$ will result in a re-distribution of population.

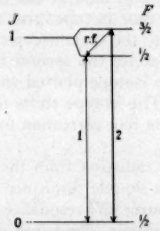


Figure 15. Transition probabilities in hyperfine structure, summed over the $2F+1$ states of each level.

Detection of the radio-frequency resonances raises another problem, since all the hyperfine states decay with equal probability. Redistribution of population changes the relative intensity of the hyperfine components, but not the total intensity. Bucka converts the relative change into an absolute change by making use of the phenomenon of self-absorption. For the stronger component is more strongly absorbed : consequently transmission of the fluorescent light through vapour of the substance concerned will discriminate between the two components. The final result will be that at radio-frequency resonance some atoms will be transferred to the levels which give rise to the more weakly absorbed light ; consequently the total amount of light which emerges from the absorbing vapour will increase.

Bucka (1958) applied the method successfully to the level $7^2P_{1/2}$ of caesium-133. The situation is rather more complicated than in our example : $J = \frac{1}{2}$ in the ground and the excited level, and $I = \frac{7}{2}$. The hyperfine interval $(F = 3)-(F = 4)$ in the ground level is greater than the Doppler width of the optical radiation $6^2S_{1/2}-7^2P_{1/2}$ (4555Å) ; that in the excited level is smaller. The line 4555Å is therefore partially resolved into two hyperfine components whose intensities are in the ratio 9 : 7.

The vessel in which the preferential self-absorption was to take place was the same in which the atoms were excited. Bucka detected a radio-frequency resonance between $F = 4$ and $F = 3$ of the level $7^2P_{1/2}$ at $400{\cdot}8 \pm 1$ Mc/s, with a signal to noise ratio of 2 or 3 : 1.

4.7. *Double Resonance Measurements of Isotope Shifts*

In a mixture of isotopes of different nuclear spin, each isotope is characterized by particular values of $\gamma_F = \omega_0/H_0$ for the various hyperfine transitions, so that by correctly choosing the magnetic field and the frequency in a double resonance experiment, a given isotope will respond independently of others of different γ_F when the mixture is suitably illuminated. All isotopes having $I = 0$ ($\gamma_F = \gamma_J$) will respond together. Suppose now that the light source provides radiation of infinitely narrow spectral range of variable, known, frequency. As the frequency is swept over the atomic resonance line one will observe a double resonance signal whose width will be the Doppler width for absorption of optical radiation in the resonance vessel. Let the magnetic field at the resonance vessel be changed to correspond to the γ_F of a different isotope, and again let the frequency of the light source be varied. Owing to the spectroscopic isotope shift, light of different frequency will be required to excite the second isotope. In this way one will obtain resonance curves for each isotope plotted on one extended frequency scale, that of the exciting radiation. The isotope shifts for the particular atomic transition may then be read off, subject to a correction for Zeeman displacements in the resonance vessel.

In an actual experiment the radiation from the light source will be Doppler-broadened, and the width of the double resonance line will be some combination of the Doppler widths of light source and resonance vessel. The double resonance method is then at a disadvantage compared with high resolution optical spectroscopy. On the other hand, the new technique is able to differentiate between isotopes in a mixture and for this reason may prove powerful in cases where the use of enriched isotopes in optical work is inconvenient or impossible.

Sagalyn, Melissinos and Bitter (1958) applied the technique to the measurement of isotope shifts in the line λ2537Å of mercury, a case in which optical measurements have not been made with enriched isotopes. A note by Melissinos (1959) has subsequently appeared, quoting measurements made on a very small sample of radioactive mercury-197. The variation in frequency of the light source was accomplished by making use of the Zeeman effect in the line λ2537Å emitted by a mercury-198 lamp. One of the σ components was isolated from the light emitted parallel to the magnetic field, variation of which up to 50 kilogauss provided a frequency range of 10^5 Mc/s (3 cm^{-1}). The isotope shifts were thus measured as a function of this magnetic field, called the ' scanning field '.

The ' splitting field ', used to select a particular isotope in the resonance vessel, was of the order of 10 kilogauss, which was sufficient to cause some departure from ' weak field ' splittings. Different transitions in one and the same isotope were therefore recorded separately since only that microwave resonance is detected for which the ' splitting field ' is correctly matched to the ' splitting frequency '. This microwave frequency was about 3000 Mc/s.

It is clear that the technique demands a preliminary study of the microwave resonance frequencies for each transition in each isotope. The line width in these preliminary experiments is that of ordinary double resonance experiments. The information gained allows the determination of g-factors and hyperfine structure constants. In the work of Sagalyn *et al.* this information is quoted for the

mercury isotopes 199 and 201 with a precision to about 15 Mc/s. The precision of the isotope shift measurements is to about 100 Mc/s.

4.8. *Detection by Coincidence of Photons in Cascade*

The following experiment proposed by Bradley (1956) is not strictly double resonance, but we include it in this section because of its close relationship.

The level $\beta(J = \frac{1}{2})$, we suppose, is populated from α with the emission of photons λ_1, and itself decays to γ with the emission of photons λ_2 (figure 16). Suppose we have detectors sensitive only to σ^+ photons of wavelengths λ_1, λ_2 respectively, and that we examine these detectors for coincidences. We expect only accidental coincidences unless radio-frequency transitions are taking place in the level β. A rise in the coincidence rate indicates radio-frequency resonance.

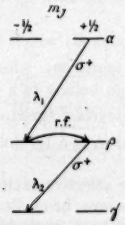


Figure 16. Photons in cascade.

Significant coincidences must occur in a time of the order of the lifetime of the level. The longer this time, the greater the possibility of accidental coincidences. Hence the signal-to-noise ratio is more favourable for levels with shorter lifetimes.

The execution of the experiment has proved to be rather difficult. No results have been published.

§ 5. Excitation by Electron Impact

As in the sections on excitation by resonance radiation, we treat the processes of excitation, radio-frequency resonance, and decay as independent processes. The two latter have already been discussed in §§ 3 and 4.3. We here briefly consider the process of excitation and the establishment of differences of population between excited states.

Let N_a be the particle density in the ground state. We suppose N_a so low that, under electron bombardment, only single encounters need be considered. Let σ_e be the cross section for excitation to a particular state, so that the rate of excitation is proportional to $N_a \sigma_e$. The equilibrium populations N_e in the excited states are determined by the equations $N_a \sigma_e \propto N_e \gamma_e$, where the γ_e are the decay constants of the excited states. Differences in the N_e clearly depend on differences in the σ_e, or in the γ_e, or in both.

We now distinguish

(*a*) states between which magnetic dipole transitions are allowed, from

(*b*) states between which electric dipole transitions are allowed.

The next three paragraphs and § 5.1 refer to class (*a*); the following paragraph and § 5.2 refer to class (*b*).

Members of class (*a*) belong to a given *term*. Under Russell–Saunders coupling they will have L and S in common, but may differ in J, or m_J, or F, or m_F. They decay at the same rate, γ_e. The establishment of differences in population therefore depends on differences in the σ_e.

It has been established in a number of cases that radiation, resulting from bombardment of atoms by a uni-directional beam of electrons whose energy is a little greater than the excitation potential of a particular term, is polarized. This implies unequal population of the states m_J or m_F which belong to that term, which is the condition required for radio-frequency resonances. Cases which are of interest to us are:

(i) Selective excitation of the states m_J which belong to a given J.

(ii) Selective excitation of the states m_F which belong to a given hyperfine level F, and of the different F which belong to a given J.

(iii) Selective excitation of the levels J (groups of m_J) of the term.

Cases (i) and (ii) are exemplified in §§ 5.1.1 and 5.1.2; case (iii) is exemplified in § 5.1.3.

The theoretical treatment of the selective excitation of different states is at the present time very inadequate. A recent, brief discussion has been given by Lamb (1957). A further contribution has been made by Baranger and Gerjuoy (1958). In an important respect (polarization at the threshold of excitation) the theoretical prediction is precisely contrary to experimental findings. Notwithstanding the theoretical uncertainty, it is helpful to know that the population disparities of different states can be quite marked (20–40%), and are functions of the energy of the electrons. That polarization should arise from a uni-directional beam of electrons whose energy is just above the threshold for excitation is not at all surprising.

Members of class (*b*) belong to different *terms*. Under Russell–Saunders coupling they generally differ in L. Their decay constants are different, and their cross sections for excitation are different. Both absolute and relative cross sections for excitation to different terms (generally summed over the states which belong to that term) have been measured and calculated in many cases. A comprehensive account, from the experimental point of view, has been given by Massey and Burhop (1952).

It is characteristic of class (*a*) that radio-frequency resonance results in a change in the polarization and spatial distribution of the decay radiation, but not in its total intensity. Resonances in class (*b*) result in changes in the total intensity of some particular *multiplet* (i.e. the totality of the transitions between two terms).

An estimate of the signal to noise ratio in a typical experiment can be made as follows (Lamb 1957):

The number of excitations per second to the level in question is $(I/e)\sigma N_a l$, where I/e is the number of electrons per second in the beam, σ is the cross section for excitation, and l is the length of vapour column of density N_a atoms/cm^3.

This quantity will also be the number of photons emitted per second, and corresponds to the quantity S at the end of § 4.3. The primary photocurrent i_s is then $I\sigma N_a l\phi\eta$, where ϕ represents the fraction of light collected by the photomultiplier, and η the efficiency of the cathode. At radio-frequency resonance, S changes by the amount ΔS.

The fraction $\beta = \Delta S/S$ is estimated as in § 3 on the basis of an assumed population distribution N_e. It is in principle independent of all the parameters mentioned in the last paragraph.

Equation (25) of § 4.3 now gives the signal to noise ratio.

5.1. *Selective Excitation of different States within a given Term*

5.1.1. *Detection by decay radiation.*

Pebay-Peyroula, Brossel and Kastler have observed magnetic dipole resonances in mercury (1957) and in sodium (private communication).

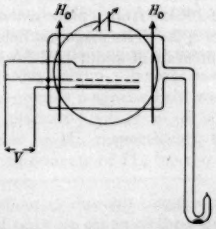


Figure 17. Excitation of mercury vapour by electron impact.

Mercury atoms at a vapour pressure determined by the temperature of the drop of mercury in the side-arm are excited in the space between the grid and plate (figure 17) by an electron beam travelling in the direction of a steady magnetic field H_0. Preferential excitation of certain states takes place, and the emitted light is polarized. Different spectrum lines are examined separately by a simple monochromator, or by filters. The excitation vessel lies within a radio-frequency coil or microwave cavity, and atomic resonances are indicated by changes in the polarization of the spectrum lines.

Experiments are usually performed at a fixed value of the frequency, and resonance curves displayed as a function of magnetic field. Interpretation of the experiments is thus a question of allocating to the atomic levels the g-values at which resonance peaks occur. This is not always a simple matter, since atomic resonances affect the polarization and intensity not only of those spectral lines which appear in primary decay, but also of those lines which appear in consequence of cascade processes. Thus, resonances in the level 7^3P_2 of mercury were detected in the optical radiation 7^3S_1–6^3P_2, 7^3S_1–6^3P_1 and 7^3S_1–6^3P_0, evidently on account of the primary decay 7^3P_2–7^3S_1. Experiments with isotopes of different nuclear spin, since they show resonances at different g_F values, assist in the interpretation.

An important detail of technique of the French group is that the distance between grid and plate is kept sufficiently short (about 4 mm) to prevent a complete cyclotron revolution of the electrons in the magnetic field H_0, otherwise very strong changes of intensity occur at the cyclotron resonance frequency.

Published measurements include the lifetime and g_J-values of the levels 5^1D_2, 6^3F_4 and 7^3P_2 in mercury, also the hyperfine structure of the level 6^3F_4 of isotopes 199 and 201.

Measurements of hyperfine structure in the levels 3^2D, 4^2D and 5^2D of sodium are being made.

5.1.2. *Detection by absorption of optical radiation.*

Dehmelt (1956) observed resonances between the states $m_J = 0, \pm 1, \pm 2$ of the level 6^3P_2 in mercury. This level is metastable, and the experiment, strictly speaking, falls outside the scope of this review. We include it, nevertheless, on account of its similarity to the other work we have described.

Excitation in mercury vapour at room temperature ($\simeq 1\mu$ vapour pressure) was by an electron current of 200 mA from a plane cathode. The states $m_J = 0, \pm 1$ were excited in preference to ± 2. The magnetic field, about 8·3 gauss, separated the energy levels by equal intervals of about 17·2 Mc/s.

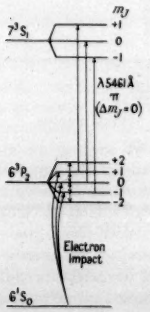


Figure 18. Detection of radio-frequency resonances in the level 6^3P_2 by their effect on the absorption of λ5461Å.

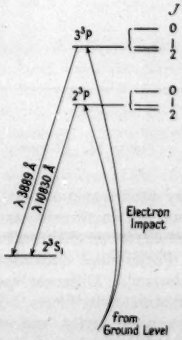


Figure 19. Excitation and decay of ^3P levels in helium.

Particular interest attaches to the method of detection of radio-frequency resonances, which was by absorption of the polarized optical radiation λ5461Å. Under the optical selection rule $\Delta m_J = 0$, the consequence of radio-frequency resonances which depletes the states $0, \pm 1$ in favour of ± 2 is a reduction in the amount of light absorbed. Since the level is metastable, only small radio-frequency fields are needed. Dehmelt shows a resonance about 0·1 gauss wide obtained under a radio-frequency field of about 12 milligauss.

Dehmelt's absorption technique, together with the 'magnetic scanning' technique of Sagalyn *et al.* (1958, § 3.7) is being applied by Bradley (private communication) to the study of resonances in the level $^3P_2$ of ^{198}Hg, ^{200}Hg, ^{202}Hg, ^{199}Hg and ^{201}Hg; in the case of the odd isotopes the hyperfine structure is seen. Natural mercury is used in the excitation vessel: the selected isotope is excited by adjusting the magnetic field at the ^{198}Hg light source to provide a Zeeman component of the correct optical frequency.

5.1.3. *Excited states of helium.*

The multiplet intervals 2^3P_2–2^3P_1, 3^3P_1–3^3P_0 and 3^3P_2–3^3P_1 in helium were measured by Lamb and Maiman (1957) and Wieder and Lamb (1957) by the techniques we have outlined.

At a pressure of up to 20 microns, helium in sealed-off tubes was excited by electron currents up to 500μA. In decay to the metastable level 2^3S_1 the levels 3^3P and 2^3P emit radiation of wavelength 3889 and 10 830Å respectively. These radiations were used to detect magnetic dipole microwave resonances in the regions of 980 Mc/s and 7000 Mc/s, and 1600 Mc/s respectively. These frequencies are not themselves the triplet intervals, but measurement of the magnetic fields at which resonances take place allows deduction of the zero-field intervals.

The excitation region of the helium tube was inserted into a resonant cavity at a place where the radio-frequency magnetic field was strong. While the cavities were designed to observe primarily either π or σ radio-frequency transitions (H_{rf} parallel or perpendicular to H_0, respectively), there were usually sufficient amounts of perpendicular components of H_{rf} for the observation of both types of transition in a given apparatus.

Detection of the 2^3P resonances through changes in the intensity in a particular direction of λ10 830Å could only be accomplished by very careful elimination of stray light from the infra-red-sensitive photomultiplier. The use of a time constant of 150 seconds in the detection system indicates the weakness of the signal and the need for careful control of all the variables of the experiment.

We have indicated that the object of this work was the accurate measurement of intervals in spectra sufficiently simple to be amenable to calculation. The resonant frequencies were determined to a precision better than 1 Mc/s. Lamb (1957) discusses the theoretical problem.

5.2. *Selective Excitation of Different Terms*

We here discuss three different experiments: that of Novick, Lipworth and Yergin (1955), also Lipworth and Novick (1957) which followed the exploratory experiment by Lamb and Skinner (1950) on $n = 2$ in singly ionized helium; that of Lamb and Sanders (1956) on $n = 3$ in hydrogen; and that of Series and Fox (1958) on $n = 4$ in singly ionized helium. The object of each of these experiments is the measurement of Lamb shifts, that is, the relative positions of the S and P levels for a given value of the principal quantum number n in a one-electron spectrum. The experiments have in common the detection of radio-frequency electric dipole transitions by 'optical' means, and the excitation of the various levels by electron impact. In each experiment differences exist in the cross section for excitation to the pair of levels concerned: more significantly, the decay

constants of the levels are different, so that the equilibrium concentrations in the excited states are different. The first two experiments use controlled excitation by a beam of electrons, and the third, a gas discharge.

5.2.1. *Singly ionized helium :* $n = 2$.

The method of detection of the radio-frequency transitions $2^2S_{1/2}$–$2^2P_{1/2}$ in this experiment is based on the entirely different probabilities for decay from these excited levels. $2^2S_{1/2}$ is metastable : its decay to the ground state of the ion, $1^2S_{1/2}$, is forbidden for electric dipole radiation. $2^2P_{1/2}$, on the other hand, decays very rapidly with the emission of radiation at 304Å (41 ev). This is detected by photoelectric emission at the platinum cathode of an Allen-type electron multiplier. An increase in the current signifies radio-frequency resonance.

The radio-frequency transition is induced in a resonant microwave cavity in which the ions have been excited from helium atoms in their ground state. The signal radiation is accompanied by a large and variable background which arises from shortlived states of the ion. This difficulty was met by Novick, Lipworth and Yergin (1955) by a pulse technique, wherein observation of the radio-frequency signal was delayed by a quarter of a microsecond after application of a pulse of exciting electrons. Lipworth and Novick (1957) used an improved method of phase detection in which, using radio-frequency amplitude modulation, they observe on two different channels the signals ' r.f. on ' and ' r.f. off '. By difference and normalization they eliminate the effect of the background, and further, by observing at two different levels of radio-frequency power, they can compensate for changes in the efficiency of the detector.

A very thorough study of the line shape and the effects of possible perturbations allows these authors to quote for the measured interval the value $14\,040{\cdot}2 \pm 4{\cdot}5$ Mc/s. The natural width of the observed resonance is about 1600 Mc/s.

5.2.2. *Hydrogen :* $n = 3$.

The level $3^2S_{1/2}$ decays more slowly than $3^2P_{1/2}$ or $3^2P_{3/2}$, and is populated more densely under electron excitation from the ground state of (molecular) hydrogen,

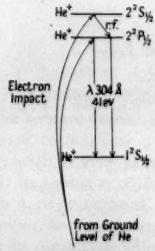


Figure 20. Excitation and decay of $2^2S_{1/2}$ and $2^2P_{1/2}$ in ionized helium.

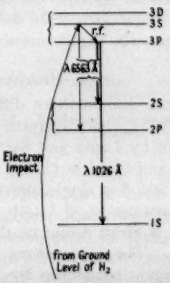


Figure 21. Excitation of 3S in atomic hydrogen, and its decay by different routes.

so that there exists the possibility of observing radio-frequency resonances between it and either of the ^{2}P levels. For our present argument, the direction of spin is immaterial ; we will refer to the S and P terms.

Both 3S and 3P radiate components of the line H_α, λ6563Å, in their decay to 2P and 3S respectively, but 3P can decay also to 1S, radiating the Lyman β line, λ1026Å. Transfer of atoms from 3S to 3P therefore changes the total intensity of the H_α complex, and this change is used to indicate radio-frequency resonance.

Lamb and Sanders (1956) used an interference filter to isolate H_α from the light emitted on bombardment by electrons of hydrogen at a pressure of about one micron. The glass envelope was situated inside a parallel plate transmission line for application of radio-frequency fields at about 3000 and 500 Mc/s. Resonances were displayed as a function of magnetic field.

A more detailed account of this experiment is in course of preparation (see Lamb and Wilcox 1958).

5.2.3. *Singly ionized helium :* $n = 4$.

The method of detection in this experiment is similar to that described above, but the ions are excited in a gas discharge rather than by controlled electron bombardment. Normal conditions in a gas discharge are such that there exists statistical equilibrium between the electrons and a set of close-lying energy levels : equality of population obtains between the atomic states. In consequence, the relative intensities of the spectral lines which derive from these states are in the ratio of the transition probabilities for spontaneous decay. In optical investigation of the components of the He II line λ4686Å, ($n = 4 \rightarrow n = 3$), Series (1954) and Herzberg (1956) noticed intensity anomalies which they ascribed to over-population of the level $4^2S_{1/2}$. Herzberg further found that the anomaly depended on the pressure of helium, and on the discharge current, and interpreted the effect in terms of the relative magnitudes of the lifetimes of the levels and the mean time between collisions with electrons. Lifetimes in excited states of ionized helium are abnormally short. If the atoms decay before they collide, statistical equilibrium will not be established.

If we accept the over-population of the level $4^2S_{1/2}$ as a fact, we are in a position to detect radio-frequency transitions to either of the 4P levels by observation of changes in intensity of the radiation 4686Å. The term diagram is essentially that of figure 21, with appropriate changes of the principal quantum numbers and of the intervals. Series and Fox (1958) are exploring the possibilities of this experiment. A monochromator is used to separate λ4686Å from the very strong radiations in the helium spectrum which accompany it, and a further interferometric resolution which will allow the partial isolation of those components affected by the radio-frequency resonance is also being attempted.

The ions in a gas discharge are clearly subject to stronger electric fields, and are therefore more likely to be perturbed than would be the case under controlled electron bombardment at low density. It is doubtful, however, whether the latter method of excitation would give sufficient intensity.

§6. Atomic Beam Experiments

In a common form of apparatus for detecting magnetic resonances in atomic beams (see, for example, Ramsey 1956), atoms in the ground states are sorted

according to their magnetic moments and therefore their m_J values in strong fields. Those atoms are detected whose m_J value changes in the experimental region, which is generally termed ' the C field '.

Changes in m_J may be brought about through the excitation of atoms by optical resonance radiation, and their subsequent decay. The time for this process is short compared with the time of flight of atoms through the C field, so that changes at the detector occur merely by illuminating the beam in the C field.

If, further, radio-frequency fields cause the excited atoms to undergo transitions, there will result further changes in the ground states, and further changes at the detector. This is the basis of the radio-frequency spectroscopy of excited atoms by atomic beam techniques.

Excited state resonances, compared with those obtained from beams of atoms in ground states, are much broader and much weaker. The breadth is of course due to the small lifetime of the excited states, which, in a typical case is about 1000 times shorter than the time of flight of an atom through the C field. The relative weakness is a consequence not only of the smaller number of excited atoms, but also of the complicated route through which changes of m_J take place.

As in experiments on ground states, the C field need not be sufficiently strong to make m_J a good quantum number. The m_J value of a state (F, m_F) for the purposes of atomic beam experiments is that of the pure state (m_J, m_I) to which (F, m_F) tends as the field is increased adiabatically. Resonances between ground states cannot normally be detected unless this value of m_J changes in the C field. In contrast, a change in m_J is not a necessary condition for the detection of resonances between excited states. For, as we have remarked, a change in m_J in the ground states may accompany the process of optical excitation and decay. Such changes may still occur when a radio-frequency transition is induced between excited states which do not differ in m_J. Radio-frequency transitions between excited states whose values of m_J are the same or different will generally result in a redistribution of the decaying atoms between the different ground states, and hence in changes in the signal.

Rabi and his colleagues have by these means measured hyperfine structures in the levels $3^2P_{3/2}$ and $^2P_{1/2}$ of ^{23}Na (Perl, Rabi and Senitzky 1955), in $5^2P_{3/2}$ and $^2P_{1/2}$ of ^{85}Rb and in $5^2P_{3/2}$ of ^{87}Rb (Senitzky and Rabi 1956), and in $4^2P_{3/2}$ and $^2P_{1/2}$ of ^{39}K (Buck and Rabi 1957), and from their measurements have deduced the nuclear electric quadrupole moments.

The first paper of this series describes the method in detail. The experiments differ in details of technique rather than of principle, and we shall not discuss them separately.

Generally, the steady magnetic field in the C region was so small that the Zeeman components belonging to a given hyperfine level F were closer than the natural width of the levels. Transitions $\Delta F = \pm 1$ were observed: the Zeeman components were quite unresolved. In each case the resonance curves were fitted to a formula of the Lorentz type, and the positions of the peaks used to determine the hyperfine structure interaction constants.

It is most valuable to have measurements on the hyperfine structures of the alkali metals by the quite independent techniques of atomic beam and double

resonance. The measurements up to the present time are largely complementary in that the atomic beam work has been restricted to the first excited states while double resonance measurements have been made mainly on second excited states (§ 4.5.2). There is no reason why double resonance measurements should not also be made on first excited states ; this has been done for sodium : the reason for the choice of second excited states is that the hyperfine structures, though smaller, are larger in relation to line width. On the other hand, resonance radiation from the second, is very much weaker than from the first excited states so that the signals are weaker. Nevertheless, the signal to noise ratio in these double resonance experiments is quite as good as that in the atomic beam experiments on first excited states. The scope of the atomic beam work could conceivably be extended by electron bombardment of the beam in the *C* field, if that were technically feasible.

§ 7. Table of Measurements

The following measurements have been made by means of the techniques described in the preceding sections :

(*a*) *g_J-values and lifetimes.*

Element	Level	g_J	($\times 10^7$ sec)	Reference
Hg	$6s6p\ ^3P_1$	$1{\cdot}4838 \pm 0{\cdot}0004$	$1{\cdot}18(\pm 0{\cdot}01)$	Brossel & Bitter (1952); Boutron *et al.* (1957)
Hg	$6s7s\ ^3S_1$	($\simeq 1{\cdot}9$)	—	Brossel & Julienne (1956)
Hg	$6s6f\ ^3F_4$	$1{\cdot}2477 \pm 0{\cdot}0004$	$4{\cdot}50 \pm 0{\cdot}12$	Pebay-Peyroula *et al.* (1957)
Hg	$6s7p\ ^3P_2$	$1{\cdot}4402 \pm 0{\cdot}0007$	$2{\cdot}10 \pm 0{\cdot}07$	Pebay-Peyroula *et al.* (1957)
Hg	$5d^9\,6s^2\,6p\ ^1D_2$	$1{\cdot}1203 \pm 0{\cdot}0003$	$16{\cdot}0 \pm 0{\cdot}16$	Pebay-Peyroula *et al.* (1957)
Na	$4p\ ^2P_{1/2}$	—	0·9†	Krüger & Scheffler (1958)
	$3p\ ^2P_{1/2}$	$1{\cdot}3341(\pm 0{\cdot}0005)$	$\simeq 0{\cdot}15$	Dodd (private communication)
K	$5p\ ^2P_{1/2}$	$1{\cdot}34 \pm 0{\cdot}02$	$\geqslant 1{\cdot}5$	Ritter & Series (1956)
Rb	$6p\ ^2P_{1/2}$	—	1·0	Meyer-Berkhout (1955)
Cs	$7p\ ^2P_{1/2}$	—	$\geqslant 1{\cdot}6$	Althoff (1955)

† The value $7{\cdot}9 \times 10^{-7}$ sec quoted in the reference is clearly a misprint.

(*b*) *Hyperfine structure coupling constants.*

(Values in brackets are either given with reserve, or have been estimated from the authors' measurements.)

Element (Isotope)	I	Level	Coupling constants (Mc/s)	$Q(\times 10^{24})$ cm^2	Reference
^{23}Na	3/2	$3p\ ^2P$	$a_{1/2}$: 19·5 ± 0·6 b : 2·4 ± 1·4	0·1 ± 0·06	Sagalyn (1954)
		$3p\ ^2P$	$a_{1/2}$: 19·06 ± 0·36 b : (2·58 ± 0·3) $a_{1/2}$: 94·45 ± 0·5	(0·100 ± 0·011)	Perl, Rabi & Senitzky (1955)
			$a_{1/2}$: 19·0(±0·5) b : 2·3(±0·3)		Dodd, J. N. (private communication)
		$4p\ ^2P$	(awaiting publication)	0·13 ± 0·04	Krüger & Scheffler (1958)
		$3, 4, 5d\ ^2D$	(awaiting publication)		Pebay *et al.* (1957)

Element (Isotope)	I	Level	Coupling constants (Mc/s)	$Q(\times 10^{24})$ cm^2	Reference
^{39}K	3/2	4p ^{2}P	$a_{1/2}$: (estimated) b : 2·8 ± 0·8 $a_{3/2}$: 28·85 ± 0·3		Buck & Rabi (1957)
		5p ^{2}P	$a_{1/2}$: 1·97 ± 0·1 b : 1·7 ± 0·3 $a_{3/2}$: (9·3 ± 0·5)	0·11 ± 0·02	Ritter & Series (1957)
^{85}Rb	5/2	5p ^{2}P	$a_{1/2}$: 25·3 ± 0·2 b : 24·4 ± 1·3 $a_{3/2}$: 120·7 ± 1·0	0·27 ± 0·02	Senitzky & Rabi (1956)
		6p ^{2}P	$a_{1/2}$: 8·16 ± 0·06 b : 8·40 ± 0·4	0·29 ± 0·02	Meyer-Berkhout (1955)
^{87}Rb	3/2	5p ^{2}P	$a_{1/2}$: 85·8 ± 0·7 b : 11·8 ± 0·6 $a_{3/2}$: 409 ± 4	0·13 ± 0·01	Senitzky & Rabi (1956)
		6p ^{2}P	$a_{1/2}$: 27·63 ± 0·1 b : 4·06 ± 0·2	0·14 ± 0·01	Meyer-Berkhout (1955)
^{133}Cs	7/2	7p ^{2}P	$a_{1/2}$: 16·60 ± 0·01 b : −0·11 ± 0·08	−0·003 ± 0·002	Althoff (1955)
			$a_{3/2}$: 100·2 ± 0·25		Bucka (1958)
^{67}Zn	5/2	4s4p ^{3}P$_1$	A : 632·3 ± 0·15 B : −19·4 ± 0·15	+0·18 ± 0·02	Böckman *et al.* (1957)
^{199}Hg	1/2	6s6p ^{3}P$_1$	A : 14910 ± 420 14733·5 ± 15		Bogle *et al.* (1957) Sagalyn *et al.* (1958)
		6s6f ^{3}F$_4$	A : 1255 ± 20		Pebay *et al.* (1957)
^{201}Hg	3/2	6s6p ^{3}P$_1$	A : −5437 ± 15 B : −35·4 ± 2·4	(0·58 ± 0·18)	Sagalyn *et al.* (1958)
		6s6f ^{3}F$_4$	A : (520 ± 25) B : (≃0)		Pebay *et al.* (1957)
^{197}Hg	1/2	6s6p ^{3}P$_1$	A : 15934 ± 30		Melissinos (1959)

(*c*) *Isotope shifts, relative to the isotope* 198, *in the line* 2537Å. ($6s^2\,{}^1S_0$–$6s\,6p\,{}^3P_1$) of mercury (Sagalyn *et al.* 1958, Melissinos 1959).

Isotope	204	202	200	198	201	201	199	197
†F (level $^3P_1$) (odd isotopes)					5/2	3/2	3/2	
Shift (Mc/s)	−15560	−10160	−4680	0	−14720	−660	+6900	+2730

The precision of measurement is to ±100 Mc/s.

† It is not necessary to specify the F value of the level $^1S_0$ since the hyperfine structure intervals are zero when $J = 0$.

(*d*) *Multiplet intervals in one- and two-electron spectra.*

Spectrum	Transition	Interval (Mc/s)	Reference
H	$3(^3P_{1/2} - {}^3S_{1/2})$	315 ± 1	Lamb & Sanders (1956)
He$^+$	$2(^3P_{1/2} - {}^3S_{1/2})$	14040·2 ± 4·5	Lipworth & Novick (1957)
He	$1s\,2p(^3P_2 - {}^3P_1)$	2291·72 ± 0·36	Wieder & Lamb (1957)
He	$1s\,3p(^3P_2 - {}^3P_1)$	658·55 ± 0·15	Wieder & Lamb (1957)
He	$1s\,3p(^3P_1 - {}^3P_0)$	8113·78 ± 0·22	Wieder & Lamb (1957)

(*e*) *Stark effect of the level* 6^3P_1 *in mercury.* (Blamont and Brossel 1956, Blamont 1957.)

For electric fields X between 10 and 70 kv cm^{-1} the difference ΔE between the displacements of the energy levels $m_J = 0$ and $m_J = \pm 1$ is given by $\Delta E = (21{\cdot}3 \pm 0{\cdot}5 \text{ Mc/s})X^2$ where X is measured in units of 10^5 v cm^{-1}.

Acknowledgments

To my colleagues in the Clarendon Laboratory I am very grateful for numerous stimulating discussions, and for their comments on the original manuscript of this paper. M. Kastler, M. Brossel and their colleagues in Paris have generously amplified in private discussions their numerous publications. For their help, and for permission to publish figures, I am most grateful. I wish also to thank Professor Krüger for permission to publish figures, and for stimulating discussions, and Professor Bradley and Dr. Dodd for making available unpublished material.

References

Althoff, K., 1955, *Z. Phys.*, **141**, 33.
Baranger, E., and Gerjuoy, E., 1958, *Proc. Phys. Soc.*, **72**, 326.
Barrat, J. P., 1957, *C.R. Acad. Sci., Paris*, **244**, 2785; 1958, *J. Phys. Radium*, **19**, 858.*
Barrat, J. P., and Winter, J., 1956, *J. Phys. Radium*, **17**, 833.
Bell, W., and Bloom, A., 1957, *Phys. Rev.*, **107**, 1559.
Besset, C., Horowitz, J., Messiah, A., and Winter, J., 1954, *J. Phys. Radium*, **15**, 251.
Blamont, J. E., 1957, *Ann. Phys., Paris*, **2**, 551.
Blamont, J. E., and Brossel, J., 1956, *C.R. Acad. Sci., Paris*, **243**, 2038.
Blamont, J. E., and Winter, J. M., 1957, *C.R. Acad. Sci., Paris*, **244**, 332.
Bloch, F., and Siegert, A., 1940, *Phys. Rev.*, **57**, 522.
Böckman, K., Krüger, H., and Recknagel, E., 1957, *Ann. Phys., Lpz.*, **20**, 250.
Bogle, G. S., Dodd, J. N., and McLean, W. L., 1957, *Proc. Phys. Soc.* B, **70**, 796.
Boutron, F., Barrat, J. P., and Brossel, J., 1957, *C.R. Acad. Sci., Paris*, **245**, 2250.
Bradley, L. C., 1956, *Phys. Rev.*, **102**, 293.
Brossel, J., and Bitter, F., 1952, *Phys. Rev.*, **86**, 308.
Brossel, J., and Julienne, C., 1956, *C.R. Acad. Sci., Paris*, **242**, 2117.
Brossel, J., and Kastler, A., 1949, *C.R. Acad. Sci., Paris*, **229**, 1213.
Brossel, J., Margerie, J., and Kastler, A., 1955, *C.R. Acad. Sci., Paris*, **241**, 865.
Brossel, J., Margerie, J., and Winter, J., 1955, *C.R. Acad. Sci., Paris*, **241**, 556, 600.
Buck, P., and Rabi, I. I., 1957, *Phys. Rev.*, **107**, 1291.
Bucka, H., 1955, *Z. Phys.*, **141**, 49; 1956, *Naturwissenschaften*, **43**, 371; 1958, *Z. Phys.*, **151**, 328.
Casimir, H., 1936, *On the Interaction between Atomic Nuclei and Electrons* (Haarlem, Holland: Teylers Tweede Genootshap).
Clendenin, W., 1954, *Phys. Rev.*, **94**, 1590.
Condon, E. U., and Shortley, G. H., 1951, *Theory of Atomic Spectra* (Cambridge: University Press).
Dayhoff, E. S., Triebwasser, S., and Lamb, W. E., 1953, *Phys. Rev.*, **89**, 106.
Dehmelt, H. G., 1956, *Phys. Rev.*, **103**, 1125; 1957, *Ibid.*, **105**, 1924.
Dirac, P. A. M., 1947, *The Principles of Quantum Mechanics*, 3rd Edn. (Oxford: University Press).
Drake, C., Lurio, A., Hughes, V. W., and White, J. A., 1958, *Bull. Amer. Phys. Soc.*, Ser. II, **3**, 7.
Gray, N. M., 1933, *Phys. Rev.*, **44**, 570.
Guiochon, M. A., Blamont, J. E., and Brossel, J., 1956, *C.R. Acad. Sci., Paris*, **243**, 1859; 1957, *J. Phys. Radium*, **18**, 99.
Güttinger, P., and Pauli, W., 1931, *Z. Phys.*, **67**, 743.
Hack, M. N., 1956, *Phys. Rev.*, **104**, 84.
Hawkins, W. B., 1955, *Phys. Rev.*, **98**, 478.

HEBERLE, J. W., REICH, H. A., and KUSCH, P., 1956, *Phys. Rev.*, **101**, 612, **104**, 1585.
HEITLER, W., 1954, *Quantum Theory of Radiation*, 3rd Edn. (Oxford : Clarendon Press).
HERZBERG, G., 1956, *Z. Phys.*, **146**, 269.
KASTLER, A., 1954, *Proc. Phys. Soc.*, A, **67**, 853 ; 1956, *J. Opt. Soc. Amer.*, **47**, 460 ; 1958, *J. Phys. Radium*, **19**, 797.*
KOPFERMANN, H., 1958, *Nuclear Moments* (New York : Academic Press Inc.) (English version by Schneider, E. E.).
KOSTER, G., 1952, *Phys. Rev.*, **86**, 148.
KRÜGER, H., and SCHEFFLER, K., 1958, *J. Phys. Radium*, **19**, 854.* A more complete paper will appear in *Z. Phys.*
KUSCH, P., 1956, *Phys. Rev.*, **101**, 627.
LAMB, W. E., 1951, *Rep. Progr. Phys.*, **14**, 23 (London : Physical Society) ; 1952, *Phys. Rev.*, **85**, 259 ; 1957, *Ibid.*, **105**, 559.
LAMB, W. E., and MAIMAN, T. H., 1957, *Phys. Rev.*, **105**, 573.
LAMB, W. E., and RETHERFORD, R. C., 1950, *Phys. Rev.*, **79**, 549 ; 1951, *Ibid.*, **81**, 222 ; 1952, *Ibid.*, **86**, 1014.
LAMB, W. E., and SANDERS, T. M., 1956, *Phys. Rev.*, **103**, 313.
LAMB, W. E., and SKINNER, M., 1950, *Phys. Rev.*, **78**, 539.
LAMB, W. E., and WILCOX, L., 1958, *J. Phys. Radium*, **19**, 801* ; *Phys. Rev.*, in the press.
LANDOLT-BÖRNSTEIN, 1950, *Zahlenwerte and Funktionen*, 6th Edn., 1/1, 260 (Berlin : Springer-Verlag).
LIPWORTH, E., and NOVICK, R., 1957, *Phys. Rev.*, **108**, 1434.
MASSEY, H. S. W., and BURHOP, E. H. S., 1952, *Electronic and Ionic Impact Phenomena* (Oxford : University Press).
MELISSINOS, A. C., 1959, *Bull. Amer. Phys. Soc.*, **4**, 11.
MEYER-BERKHOUT, U., 1955, *Z. Phys.*, **141**, 185.
MITCHELL, A. C. G., and ZEMANSKY, M. W., 1934, *Resonance Radiation and Excited Atoms* (Cambridge : University Press).
NOVICK, R., and COMMINS, E., 1958, *Phys. Rev.*, **111**, 822.
NOVICK, R., LIPWORTH, E., and YERGIN, P. F., 1955, *Phys. Rev.*, **100**, 1153.
PEBAY-PEYROULA, J.-C., BROSSEL, J., and KASTLER, A., 1957, *C.R. Acad. Sci., Paris*, **244**, 57, **245**, 840.
PERL, M. L., RABI, I. I., and SENITZKY, B., 1955, *Phys. Rev.*, **98**, 611.
PHILLIPS, M., 1956, *Phys. Rev.*, **103**, 322.
PRINGSHEIM, P., 1949, *Fluorescence and Phosphorescence* (London : Interscience Publishers).
RABI, I. I., 1937, *Phys. Rev.*, **51**, 652.
RAMSEY, N. F., 1955, *Phys. Rev.*, **100**, 1191 ; 1956, *Molecular Beams* (Oxford : University Press).
RITTER, G. J., and SERIES, G. W., 1957, *Proc. Roy. Soc.* A, **238**, 473.
SAGALYN, P. L., 1954, *Phys. Rev.*, **94**, 885.
SAGALYN, P. L., MELISSINOS, A. C., and BITTER, F., 1958, *Phys. Rev.*, **109**, 375.
SALWEN, H., 1955, *Phys. Rev.*, **99**, 1274.
SENITZKY, B., and RABI, I. I., 1956, *Phys. Rev.*, **103**, 315.
SERBER, R., 1956, *Special Technical Report of Columbia Radiation Laboratory.*
SERIES, G. W., 1954, *Proc. Roy. Soc.* A, **226**, 377 ; 1957, *Spectrum of Atomic Hydrogen* (Oxford : University Press).
SERIES, G. W., and FOX, W. N., 1958, *J. Phys. Radium*, **19**, 850.*
STERNHEIMER, R. M., 1954, *Phys. Rev.*, **95**, 736 and earlier papers there cited.
TORREY, H. C., 1941, *Phys. Rev.*, **59**, 293.
TRIEBWASSER, S., DAYHOFF, E. S., and LAMB, W. E., 1953, *Phys. Rev.*, **89**, 98.
WEISSKOPF, V. F., and WIGNER, E. P., 1930, *Z. Phys.*, **63**, 54, **65**, 18.
WESEMEYER, H., and DANIELS, J., 1958, *Canad. J. Phys.*, **36**, 405.
WIEDER, I., and LAMB, W. E., 1957, *Phys. Rev.*, **107**, 125.
WINTER, J., 1955, *C.R. Acad. Sci., Paris*, **241**, 375, 600.

* These papers were read at the Magnetic Resonance Conference in Paris in July 1958.

THE THEORY OF A FERMI LIQUID

(The Properties of Liquid ^{3}He at Low Temperatures)

By A. A. ABRIKOSOV and I. M. KHALATNIKOV
Institute for Physical Problems, Moscow

Translated by M. G. Priestley

CONTENTS

	PAGE
§ 1. Introduction	330
§ 2. The energy of excitations	331
§ 3. Effective mass	332
§ 4. Specific heat and entropy	333
§ 5. Magnetic susceptibility	334
§ 6. Kinetic equation	336
§ 7. Viscosity	337
§ 8. Thermal conductivity	341
§ 9. Sound	343
§ 10. Dispersion and absorption of sound	348
§ 11. Light scattering (fluctuations in the distribution function)	352
Appendix: The microscopic theory of a Fermi liquid	358
1. A rarefied Fermi gas	358
2. The microscopic theory of a Fermi liquid at $T = 0$	362
References	366

Abstract. The review gives the general theory of an isotropic Fermi liquid constructed by L. D. Landau, and continued by the authors of the present article. The results of the theory are compared with the experimental data on the properties of liquid helium 3 at low temperatures.

The contents of the review are set out in the following order. After a short introduction (§ 1), § 2 gives the basic premises of L. D. Landau's theory, and in particular the energy of excitations and the interaction function f are derived. § 3 gives the derivation from Galileo's principle of a relation which expresses the effective mass of the excitations in terms of the mass of the atoms and the zeroth harmonic of the function f.

§ 4 is devoted to the derivation of the specific heat of a Fermi liquid, and an expression for the magnetic susceptibility is deduced in § 5. The exchange interaction of the excitations plays a considerable role in this expression. § 6 gives the kinetic equation for the excitations and this is used to find expressions for the momentum and energy flux. The results of § 6 are applied in § 7 to the calculation of the coefficient of viscosity, and in § 8 the coefficient of thermal conductivity is found by an analogous method. § 9 is concerned with the problem of the velocity of sound. In particular, this section contains the result obtained by L. D. Landau concerning the possibility of propagating the so-called 'zero sound' in a Fermi liquid at sufficiently low temperatures. § 10 considers the problem of dispersion and absorption of sound. § 11 deduces the average fluctuations of the distribution function and this result is used to calculate the scattering of light. An expression is given for the total intensity of scattered light, and the frequency and angular distributions are also given.

The Appendix shows how the basic premises of the theory of a Fermi liquid can be derived from microscopic considerations.

§ A1 considers the properties of a rarefied Fermi gas. § A2 gives the microscopic basis of the energy spectrum of a Fermi liquid in the general case, and also gives a relation between the function f and the scattering amplitude of excitations.

§ 1. Introduction

The investigation of the superfluidity of helium II stated the question of the properties of a quantum liquid, i.e. of a system of interacting particles whose behaviour is largely conditioned by quantum effects. L. D. Landau's theory of superfluidity was at the same time the first theory of a quantum liquid. It described the properties of so-called Bose liquids, i.e. of quantum systems of interacting particles, the excitations of which obey Bose statistics.

However, besides these liquids there are others which form a much wider class—the so-called Fermi liquids, the excitations of which have a spin $\frac{1}{2}$ and obey Fermi statistics. This class contains liquid helium 3, electrons in metals and possibly heavy nuclei, although of course the last case is quite specific. Until very recently there was no real theory of a Fermi liquid, and theoretical calculations were usually limited to an approach analogous to that made in Bose liquids : some energy spectrum was postulated and then substituted in the Fermi distribution formula.

In view of the fact that in metals the problem was considerably complicated by the anisotropy of the spectra, it was difficult to test the adequacy of this approach. However, in the case of liquid helium 3 at low temperatures it became quite clear that such an approach is not always successful.

The consistent theory of a Fermi liquid was developed by Landau (1956). He showed that, contrary to Bose liquids, the interaction of the excitations plays a considerable role in Fermi liquids, and hence it is not always possible to consider the excitations in a Fermi liquid as an ideal gas.

In this paper we describe the theory of a Fermi liquid as applied to an isotropic model and we show how the Landau theory enables us to explain various properties of liquid helium 3. In addition to this we give an account of the recently completed work of Landau, which shows how the basic assumptions of the theory of a Fermi liquid follow from a microscopic consideration of the interaction ; we also give an account of the theory of a rarefied Fermi gas. Here it is possible to obtain the fundamental quantities of the theory of a Fermi liquid by direct calculation.

Before going on to an account of the theory, and for convenience in its later applications, we give here the main properties of liquid helium 3. The isotope of helium with atomic weight 3 liquefies at a temperature of 3·2°K (critical point $T_c = 3{\cdot}3°, p_c = 845$ mm Hg) and at pressures of less than 30 atmospheres remains liquid right down to the lowest temperatures at which it has been investigated ($\sim 0{\cdot}2°$K). This happens for the same reasons as in helium II—the weakness of the interaction between the atoms and their small mass. This last point leads to the fact that at low temperatures the de Broglie wavelength corresponding to the motion of atoms of helium 3 can be greater than the interatomic distance, i.e. a quantum liquid results.

Right down to the lowest temperatures investigated, liquid helium 3 does not become superfluid. In view of the fact that according to the theory of Landau every Bose liquid must be superfluid, while Fermi liquids do not have this property, helium 3 should be considered as a Fermi liquid. It should be pointed out that this circumstance is not trivial, and does not follow automatically from the fact that the atoms of helium 3 have a spin of $\frac{1}{2}$. A system of such atoms

could have exclusively Bose excitations and furthermore, as will be shown later, such excitations do actually exist, although they do not play an important role in the determination of the properties of helium 3. Only the absence of superfluidity enables us to conclude with confidence that helium 3 is a Fermi liquid.

§ 2. The Energy of Excitations

The possibility of describing an excited state of liquid helium 3 in terms of a gas of quasi-particles with a corresponding energy distribution function is based on the fact that the interaction of the particles of a Fermi gas falls off quickly as the temperature is reduced. From general considerations it is clear that the time between collisions is proportional to the square of the smearing out of the Fermi distribution, i.e. on the average to the square of the temperature. At the same time the average energy (taken as the distance from the Fermi boundary) of Fermi particles in the zone of smearing out, i.e. those that determine the macroscopic properties of the system, is directly proportional to the temperature. Because of this the 'indeterminacy of energy' due to collisions will be small compared with the average energy itself at a sufficiently low temperature. It may also be in the terms that at sufficiently low temperatures the damping of excitations, described by the imaginary part of their energy, will be small compared with the real part of the energy.

There are two fundamental assumptions which form the basis of Landau's theory. The first of these is that the classification of the levels of a Fermi liquid corresponds to the classification of the levels of non-interacting atoms. This means that for a gradual 'switching on' of the interaction the atoms gradually become 'quasi-particles', each of which has a definite energy. Thus the number of quasi-particles equals the number of atoms.

The second assumption is that the interaction of the quasi-particles can be taken into account as a self-consistent field of surrounding particles, which leads to the fact that the energy of the system does not equal the sum of the energies of the quasi-particles, but is a functional of their distribution function. The energy of a quasi-particle must be determined as a variational derivative of the energy density in momentum space, i.e.

$$\delta E = \int \epsilon \delta n \, d\tau \qquad \ldots\ldots (2.1)$$

where $d\tau = 2dp_x\,dp_y\,dp_z/(2\pi\hbar)^3$. The factor 2 in $d\tau$ takes into account the spin $\frac{1}{2}$ of the quasi-particle.

In some cases it is necessary to take into account the dependence of the energy of the particles on their spin. Since the spin is a quantum mechanical quantity, the distribution function must here be considered as a statistical matrix, and instead of (2.1) we use a definition of ϵ in the form

$$\delta E = \tfrac{1}{2}\mathrm{Sp}_\sigma \int \epsilon \delta n \, d\tau. \qquad \ldots\ldots (2.2)$$

Definition of the energy of a particle by (2.2) leads to the conclusion that the

equilibrium distribution functions are Fermi functions. To prove this it is most convenient to use the expression for the entropy†

$$S = -\tfrac{1}{2}\mathrm{Sp}_\sigma \int \{n \ln n + (1-n)\ln(1-n)\}\, d\tau \qquad \ldots\ldots(2.3)$$

as this equation has a purely combinatorial origin. From the condition for maximum entropy, keeping the number of particles and the energy constant,

$$N = \tfrac{1}{2}\mathrm{Sp}_\sigma \int n d\tau = \text{const.}, \quad E = \text{const.}$$

we can find the distribution function by variation with respect to δn

$$n(\epsilon) = n_{\mathrm{F}}(\epsilon) = \frac{1}{\exp\{(\epsilon-\mu)/T\}+1}. \qquad \ldots\ldots(2.4)$$

Since it is a functional of n, the energy ϵ depends on the temperature. This dependence can be expressed in the following form. If we write the equilibrium energy of the quasi-particles at $T=0$ as $\epsilon_0(\mathbf{p},\boldsymbol{\sigma})$, then for small deviations from equilibrium or for small $T>0$ it will be expressed by the equation

$$\begin{aligned}\epsilon &= \epsilon_0(\mathbf{p},\boldsymbol{\sigma}) + \delta\epsilon(\mathbf{p},\boldsymbol{\sigma}) \\ &= \epsilon_0(\mathbf{p},\boldsymbol{\sigma}) + \tfrac{1}{2}\mathrm{Sp}_{\sigma'}\int f(\mathbf{p},\boldsymbol{\sigma};\,\mathbf{p}',\boldsymbol{\sigma}')\,\delta n(\mathbf{p}',\boldsymbol{\sigma}')\, d\tau'. \qquad \ldots\ldots(2.5)\end{aligned}$$

Here $\delta n = n - n_{\mathrm{F}}$ (0), and f is an operator which depends on the momenta and spin operators of two particles. Because of its definition as a second variational derivative of E with respect to δn, the function f must be symmetrical with respect to interchange of $\mathbf{p},\boldsymbol{\sigma}$ and $\mathbf{p}',\boldsymbol{\sigma}'$. The function f is a very important characteristic of a Fermi liquid. It was shown by Landau that f is related to the forward scattering amplitude of two quasi-particles (see Appendix § A2).

§ 3. Effective Mass

Since we are considering low temperatures, in the absence of an external field ϵ_0 may be written as

$$\epsilon_0 - \mu(0) = v(p - p_0) \qquad \ldots\ldots(3.1)$$

where p_0 and v are respectively the momentum and velocity at the Fermi surface. The velocity v can be written as

$$v = p_0/m^* \qquad \ldots\ldots(3.2)$$

where m^* is the effective mass. As was shown by Landau, there is a definite relation between m^* and f which follows from Galileo's principle of relativity.

According to Galileo's principle, the momentum per unit volume must be equal to the flux of mass of the liquid. Since the velocity of a quasi-particle can be defined as the derivative $\partial\epsilon/\partial\mathbf{p}$, and their number equals that of atoms, we find

$$\tfrac{1}{2}\mathrm{Sp}_\sigma \int \mathbf{p} n\, d\tau = \tfrac{1}{2}\mathrm{Sp}_\sigma \int m \frac{\partial\epsilon}{\partial\mathbf{p}} n d\tau \qquad \ldots\ldots(3.3)$$

† Henceforth we shall use the energy units for temperature, i.e. $k = 1$.

where m is the mass of an atom. We take the variational derivative of this expression with respect to n

$$\tfrac{1}{2}\mathrm{Sp}_\sigma \int \mathbf{p}\,\delta n\,d\tau = \tfrac{1}{2}\mathrm{Sp}_\sigma \int m\frac{\partial \epsilon}{\partial \mathbf{p}}\,\delta n\,d\tau + \tfrac{1}{2}\mathrm{Sp}_\sigma \mathrm{Sp}_{\sigma'} \iint mn\frac{\partial}{\partial \mathbf{p}} f(\mathbf{p},\sigma;\,\mathbf{p}',\sigma')\,\delta \mathbf{n}'\,d\tau\,d\tau'.$$

Exchanging $\mathbf{p},\sigma$ and $\mathbf{p}',\sigma'$ in the last term and using the fact that δn is arbitrary, we find

$$\frac{\mathbf{p}}{m} = \frac{\partial \epsilon}{\partial \mathbf{p}} + \tfrac{1}{2}\mathrm{Sp}_{\sigma'} \int \frac{\partial f}{\partial \mathbf{p}} n'\,d\tau' = \frac{\partial \epsilon}{\partial \mathbf{p}} - \tfrac{1}{2}\mathrm{Sp}_{\sigma'} \int f\frac{\partial n}{\partial \mathbf{p}}\,d\tau. \qquad \ldots\ldots(3.4)$$

If the system is not in an external magnetic field, the relation (3.4) does not lose generality if $\frac{1}{2}\mathrm{Sp}_\sigma$ is taken. In the equilibrium case at $T = 0$, ϵ may be taken in the form (3.1) and $\partial n/\partial \mathbf{p}$ replaced by $-\delta(p - p_0)$ in (3.4). Thus we find

$$\frac{1}{m} = \frac{1}{m^*} + \frac{p_0}{2(2\pi\hbar)^3}\,\mathrm{Sp}_\sigma \mathrm{Sp}_{\sigma'} \int f(\chi)\cos\chi\,d\Omega \qquad \ldots\ldots(3.5)$$

where $f(\chi)$ is the value of the function f when $|\mathbf{p}| = |\mathbf{p}'| = p_0$. It is natural that apart from negligible corrections this relation is conserved for temperatures near $T = 0$ and for small deviations from equilibrium.

§ 4. Specific Heat and Entropy

Here we will assume that there is no external magnetic field, so that the quantities will not depend on spin.

A knowledge of the energy spectrum makes it possible to determine the specific heat of a Fermi liquid (see Abrikosov and Khalatnikov 1957 a). It can be found in the usual way by differentiating the energy with respect to the temperature for a given number of particles.

Here it is important that the change in energy density is expressed by (2.1), and hence we can write

$$C = \left(\frac{\partial E}{\partial T}\right)_N = \int \epsilon \left(\frac{\partial n}{\partial T}\right)_N d\tau \simeq \int \epsilon_0 \left(\frac{\partial n}{\partial T}\right)_N d\tau. \qquad \ldots\ldots(4.1)$$

The replacement of ϵ by ϵ_0 is justified by the fact that while the main term in the specific heat is linear in T, the correction in (2.5) gives only cubic terms.

Indeed δn can be written in the form

$$\delta n = n(T) - n(0) = \int_0^T \frac{\partial n}{\partial T}\,dT = \int_0^T \frac{\partial n}{\partial \epsilon}\left(-\frac{\epsilon - \mu}{T} + \frac{\partial \epsilon}{\partial T} - \frac{\partial \mu}{\partial T}\right) dT.$$

At low temperatures the derivative of the Fermi function has the form

$$\frac{\partial n}{\partial \epsilon} \simeq -\delta(\epsilon - \mu) - \frac{\pi^2}{6}\,T^2 \frac{\partial^2 \delta(\epsilon - \mu)}{\partial \epsilon^2}.$$

Using this,

$$\delta\epsilon = \int f\,\delta n'\,d\tau' = \frac{\pi^2}{6}\,T^2 \left[\frac{\partial}{\partial \epsilon'}\left(f\frac{dT'}{d\epsilon'}\right)\right]_{\epsilon'=\epsilon} - \int_0^T \left[f\left(\frac{\partial \epsilon'}{\partial T} - \frac{\partial \mu}{\partial T}\right)\frac{dT'}{d\epsilon'}\right]_{\epsilon'=\mu} dT. \qquad \ldots\ldots(4.2)$$

The derivative $\partial\mu/\partial T$ equals the entropy which is, as we shall see, linear in T. The order of the term $\partial\epsilon'/\partial T$ may be established by using (4.2), which, on differentiation with respect to T, easily gives that $\partial\epsilon/\partial T$ is first order in T, and consequently $\delta\epsilon$ is second order.

Because of the replacement of ϵ by ϵ_0 the energy can be simply calculated as the integral

$$E = \int \epsilon_0 n \, d\tau \quad \text{for} \quad N = \int n d\tau = \text{const.}$$

and again, with accuracy to cubic terms in T, ϵ may be replaced by ϵ_0 in the functions n which appear in the integrals for E and N. After this the calculation of the integrals does not differ from the usual case of a Fermi gas.

Thus we find

$$c = \frac{C}{N} = \gamma T, \quad \gamma = \frac{\pi^2}{3N}\left(\frac{d\tau}{d\epsilon}\right)_\mu = \left(\frac{\pi}{3N}\right)^{2/3} \frac{m^*}{\hbar}. \qquad \ldots\ldots(4.3)$$

Because of the linear temperature dependence, the entropy is equal to the specific heat.

By comparing (4.3) with the experimental data on the entropy of liquid helium 3 at low temperatures we can find the parameter γ, and consequently determine the effective mass. Unfortunately, the experimental curve has as yet only reached the beginning of the linear part (Abraham, Osborne and Weinstock 1955); however, an approximate value of γ can be obtained from these data, which is about 3 cal $\text{mol}^{-1}\,\text{deg}^{-2}$. Since the number of quasi-particles equals the number of atoms, we can also determine the momentum at the Fermi surface, knowing the density of liquid helium 3 ($\rho = 0{\cdot}078\ \text{g cm}^{-3}$). Thus we find

$$m^* = 1{\cdot}43m\ (^3\text{He}), \quad p_0/\hbar = 0{\cdot}76 \times 10^8\ \text{cm}^{-1}. \qquad \ldots\ldots(4.4)$$

§ 5. Magnetic Susceptibility

Since particles which have spin must have a magnetic moment, a Fermi liquid must be magnetic. In the isotropic case there are only two possibilities: the liquid can either be paramagnetic or ferromagnetic.

Anti-ferromagnetism is not possible in an isotropic Fermi liquid. Here we give an account only of a paramagnetic Fermi liquid, as liquid ^{3}He is of this type. A calculation of the susceptibility of such a liquid was given by Landau (1956, see also Abrikosov and Khalatnikov 1957 a).

When a system is placed in a magnetic field $\mathbf{H}$ the particles acquire an additional energy which depends on their spin. If they were free this change in energy would be $-\beta(\boldsymbol{\sigma}\mathbf{H})$. However, in a Fermi liquid it is also necessary to take into account the fact that the distribution function is altered. Thus we get

$$\delta\epsilon = -\beta(\boldsymbol{\sigma}\mathbf{H}) + \tfrac{1}{2}\text{Sp}_{\sigma'}\int f \, \delta n' d\tau'. \qquad \ldots\ldots(5.1)$$

Here the dependence of the function f on the spins is important. In the general case f can be written as

$$f(\mathbf{p}, \boldsymbol{\sigma}; \mathbf{p}', \boldsymbol{\sigma}') = f(\mathbf{p}, \mathbf{p}') + \zeta_{ik}(\mathbf{p}, \mathbf{p}')\,\sigma_i \sigma_k'. \qquad \ldots\ldots(5.2)$$

However, it is natural to assume that the spin interaction has essentially a purely exchange origin. Hence f becomes

$$f(\mathbf{p}, \boldsymbol{\sigma}; \mathbf{p}', \boldsymbol{\sigma}') = f(\mathbf{p}, \mathbf{p}') + \zeta(\mathbf{p}, \mathbf{p}')\,\boldsymbol{\sigma}\boldsymbol{\sigma}'. \qquad \ldots\ldots(5.3)$$

The change in the distribution function δn in (5.1) is related only to the change in energy ϵ, since the change in the chemical potential is only second order in H. Hence we may substitute $\delta n = (\partial n/\partial \epsilon)\,\delta\epsilon$ in (5.1). This gives

$$\delta\epsilon = -\beta(\boldsymbol{\sigma}\mathbf{H}) + \tfrac{1}{2}\mathrm{Sp}_{\sigma'}\int f \frac{\partial n'}{\partial \epsilon'}\,\delta\epsilon'\,d\tau' \qquad \ldots\ldots(5.4)$$

as the equation for $\delta\epsilon$. This quantity may be assumed to be of the form

$$\delta\epsilon = -\xi(\boldsymbol{\sigma}\mathbf{H}) \qquad \ldots\ldots(5.5)$$

and, substituting in (5.4), we get the equation

$$\xi = \beta + \tfrac{1}{4}\int \zeta \frac{\partial n'}{\partial \epsilon'}\,\xi\,d\tau'. \qquad \ldots\ldots(5.6)$$

Since $\partial n'/\partial\epsilon' \simeq -\delta(\epsilon' - \mu)$, the integration is carried out over the Fermi surface. Below we shall see that only the values of ξ taken at the Fermi surface are important in the magnetic susceptibility. In this case both the arguments of $\zeta(\mathbf{p}, \mathbf{p}')$ have modulus p_0 and ζ depends only on the angle between them. Denoting $\int\zeta(\theta)\,d\Omega/4\pi$ by $\bar{\zeta}$, we get from (5.6)

$$\xi = \frac{\beta}{1 + \tfrac{1}{4}\bar{\zeta}(d\tau/d\epsilon)_\mu}. \qquad \ldots\ldots(5.7)$$

The susceptibility is given by the formula

$$\chi = \frac{\partial}{\partial \mathbf{H}}\tfrac{1}{2}\mathrm{Sp}\int n\,\beta\boldsymbol{\sigma}\,d\tau.$$

Only the spin-dependent part of n is important in this expression. Thus we find

$$\chi = \tfrac{1}{4}\beta\xi\left(\frac{d\tau}{d\epsilon}\right)_\mu = \frac{\tfrac{1}{4}\beta^2(d\tau/d\epsilon)_\mu}{1 + \tfrac{1}{4}\bar{\zeta}(d\tau/d\epsilon)_\mu}. \qquad \ldots\ldots(5.8)$$

The quantity $(d\tau/d\epsilon)_\mu$ can be expressed in terms of the coefficient γ in the linear specific heat law

$$\frac{1}{\chi} = \beta^{-2}\left(\bar{\zeta} + \frac{4\pi^2}{3\gamma N}\right). \qquad \ldots\ldots(5.9)$$

The expression (5.9) can be compared with the experimental data on the magnetic susceptibility of ^{3}He (Fairbank, Ard and Walters 1954, see also Abrikosov and Khalatnikov 1957 a). These show that $\bar{\zeta}$ is negative and of magnitude approximately 0·85 of the second term. Thus the exchange interaction has a substantial effect on the magnitude of the magnetic susceptibility of liquid helium 3. The sign of the effect is such that a parallel spin orientation is favoured, but this does not lead to ferromagnetism because the Fermi tendency towards antiparallel spins is stronger. It is possible that increase of pressure could change the parameters in such a way that ferromagnetism would be possible. However, there are as yet no experimental data in favour of this conclusion.

§6. Kinetic Equation

In the absence of a magnetic field the kinetic equation for the distribution function has the usual form

$$\frac{\partial n}{\partial t}+\frac{\partial n}{\partial \mathbf{r}}\frac{\partial \epsilon}{\partial \mathbf{p}}-\frac{\partial n}{\partial \mathbf{p}}\frac{\partial \epsilon}{\partial \mathbf{r}}=I(n) \qquad \ldots\ldots(6.1)$$

where $I(n)$ is the collision integral. Here it should be noted that the energy ϵ is a functional of the distribution function and thus also depends on the coordinates. From this kinetic equation, Landau (1956) obtained expressions for the energy and momentum flux.

To find the momentum flux (6.1) must be multiplied by p_i and integrated over the relevant phase volume. As a result of conservation of momentum in collisions the integral $\int p_i I(n)\,d\tau$ is zero and we are left with the equation

$$\frac{\partial}{\partial t}\int p_i n\,d\tau+\int p_i\left(\frac{\partial n}{\partial x_k}\frac{\partial \epsilon}{\partial p_k}-\frac{\partial n}{\partial p_k}\frac{\partial \epsilon}{\partial x_k}\right)d\tau=0.$$

The expression in brackets in the second integral can be written as

$$\frac{\partial}{\partial x_k}\left(n\frac{\partial \epsilon}{\partial p_k}\right)-\frac{\partial}{\partial p_k}\left(n\frac{\partial \epsilon}{\partial x_k}\right).$$

Substituting this in the integral and integrating the second term by parts, we get

$$\frac{\partial}{\partial x_k}\int p_i\frac{\partial \epsilon}{\partial p_k}n\,d\tau+\int n\frac{\partial \epsilon}{\partial x_i}d\tau.$$

The last term may be written as

$$\frac{\partial}{\partial x_i}\int n\epsilon\,d\tau-\int \epsilon\frac{\partial n}{\partial x_i}d\tau,$$

but since $\delta E=\int \epsilon\,\delta n\,d\tau$, the last term is $\partial E/\partial x_i$. Thus we obtain the equation

$$\frac{\partial}{\partial t}\int p_i n\,d\tau+\frac{\partial \Pi_{ik}}{\partial x_k}=0, \qquad \ldots\ldots(6.2)$$

which expresses the law of conservation of momentum. The tensor Π_{ik} is the momentum flux and is given by

$$\Pi_{ik}=\int p_i\frac{\partial \epsilon}{\partial p_k}n\,d\tau+\delta_{ik}\left[\mathrm{Sp}\int n\epsilon\,d\tau-E\right]. \qquad \ldots\ldots(6.3)$$

The law of conservation of energy is obtained in an analogous manner. The kinetic equation is multiplied by ϵ and integrated with respect to $d\tau$. Because energy is conserved in the collisions, the integral $\int \epsilon I(n)\,d\tau$ is zero and we find

$$\int \epsilon\frac{\partial n}{\partial t}d\tau+\int \epsilon\left(\frac{\partial n}{\partial \mathbf{r}}\frac{\partial \epsilon}{\partial \mathbf{p}}-\frac{\partial n}{\partial \mathbf{p}}\frac{\partial \epsilon}{\partial \mathbf{r}}\right)d\tau=0.$$

Writing the first term in brackets in the form

$$\frac{\partial}{\partial \mathbf{r}}\int n\epsilon\frac{\partial \epsilon}{\partial \mathbf{p}}d\tau-\int n\frac{\partial \epsilon}{\partial \mathbf{r}}\frac{\partial \epsilon}{\partial \mathbf{p}}d\tau-\int n\frac{\partial^2 \epsilon}{\partial \mathbf{r}\,\partial \mathbf{p}}\epsilon\,d\tau$$

and integrating the second term in brackets by parts with respect to $\mathbf{p}$, we obtain the law of conservation of energy

$$\int \epsilon \frac{\partial n}{\partial t} d\tau + \operatorname{div} \mathbf{Q} = 0 \qquad \ldots\ldots(6.4)$$

where the energy flux is

$$\mathbf{Q} = \int n\epsilon \frac{\partial \epsilon}{\partial \mathbf{p}} d\tau. \qquad \ldots\ldots(6.5)$$

§ 7. Viscosity

Using (6.1) and the expressions for the fluxes of energy (6.5) and momentum (6.2), we can find the coefficients of viscosity and the thermal conductivity of a Fermi liquid (Abrikosov and Khalatnikov 1957 b).

First we shall find the viscosity. To do this we imagine that there is in the liquid a motion with a spatially slightly inhomogeneous velocity $\mathbf{u}$. The distribution function will be only slightly different from its equilibrium value

$$n = n_0 + \delta n \qquad \ldots\ldots(7.1)$$

where

$$n_0 = \frac{1}{\exp\{(\epsilon_0 - \mathbf{pu} - \mu)/T\} + 1}; \quad |\delta n| \ll n_0 \qquad \ldots\ldots(7.2)$$

and δn is found from the kinetic equation (6.1). As usual it is necessary to substitute the function n_0 in the left-hand side of the kinetic equation. We shall suppose that $\mathbf{u} = 0$ at the point considered and by substituting (7.2) in (6.1) we find

$$\frac{\partial n_0}{\partial t} - \frac{1}{3}\frac{\partial n_0}{\partial \epsilon}\mathbf{p}\frac{\partial \epsilon}{\partial \mathbf{p}} \operatorname{div} \mathbf{u} - \frac{1}{2}\frac{\partial n_0}{\partial \epsilon}\left(p_i \frac{\partial \epsilon}{\partial p_k} - \frac{1}{3} p_l \frac{\partial \epsilon}{\partial p_l}\delta_{ik}\right) \times \left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} - \frac{2}{3}\delta_{ik}\frac{\partial u_l}{\partial x_l}\right) = I(n). \qquad \ldots\ldots(7.3)$$

We shall now transform $\partial n_0/\partial t$ and shall show that this expression is also independent of the term in f in (2.5). Using (7.2) we can write

$$\delta n_0 = \frac{\partial n_0}{\partial \epsilon_0}\left(\delta\epsilon_0 - \delta\mu - \frac{\epsilon_0 - \mu}{T}\delta T - \mathbf{p}\delta\mathbf{u}\right), \qquad \ldots\ldots(7.4)$$

but since the derivative $\partial n_0/\partial \epsilon_0$ is non-zero only in a small range about $\epsilon_0 = \mu$ and is also quickly varying over this range, we may consider that the quantities in the bracket refer to this point (the correction will be relatively of the order of $(T/\mu)^2$). Here the changes $\delta\epsilon_0$ and $\delta\mu$ will be arbitrary and will not be equal. On the other hand, the distribution function is normalized by the relation $\int n_0 d\tau = N$ where N is the number of atoms per unit volume. By variation of this relation we find

$$\delta N = \int \frac{\partial n_0}{\partial \epsilon_0}\left(\delta\epsilon_0 - \delta\mu - \frac{\epsilon_0 - \mu}{T}\delta T - \mathbf{p}\delta\mathbf{u}\right)$$

$$\simeq -\left[(\delta\epsilon_0 - \delta\mu - \mathbf{p}\delta\mathbf{u})\frac{d\tau}{d\epsilon_0}\right]_{\epsilon_0 = \mu} \qquad \ldots\ldots(7.5)$$

since $\int(\partial n_0/\partial\epsilon_0)\,d\epsilon_0 = -1$. By comparing (7.5) with (7.4) we get

$$\frac{\partial n_0}{\partial t} = -\frac{\partial N}{\partial t}\frac{\partial n_0}{\partial \epsilon}\left(\frac{d\epsilon_0}{d\tau}\right)_{\epsilon_0=\mu}.$$

The number of atoms N satisfies the continuity condition

$$\frac{\partial N}{\partial t} + N \operatorname{div}\mathbf{u} = 0. \qquad \ldots\ldots(7.6)$$

Thus the term in $\partial n_0/\partial t$ gives an extra $\operatorname{div}\mathbf{u}$, and as a result the whole term in $\operatorname{div}\mathbf{u}$ becomes

$$-\frac{\partial n_0}{\partial \epsilon}\left[\frac{1}{3}\mathbf{p}\frac{\partial \epsilon_0}{\partial \mathbf{p}} - N\left(\frac{d\epsilon_0}{d\tau}\right)_{\epsilon_0=\mu}\right]\operatorname{div}\mathbf{u}. \qquad \ldots\ldots(7.7)$$

Substituting (3.1) and using the equality of the number of excitations and the number of particles, as a consequence of which $p_0/\hbar = (3\pi^2 N)^{1/3}$, we find that this expression reduces to zero. Thus neglecting small terms of the order of about $(T/\mu)^2$ there are no terms in $\operatorname{div}\mathbf{u}$ on the left-hand side of the kinetic equation.

On the right-hand side of (6.1) we have the collision integral

$$I(n) = -\int w[n_1 n_2(1-n_1')(1-n_2') - (1-n_1)(1-n_2)n_1' n_2']$$
$$\times\,\delta(\mathbf{p}_1+\mathbf{p}_2-\mathbf{p}_1'-\mathbf{p}_2')\,\delta(\epsilon_1+\epsilon_2-\epsilon_1'-\epsilon_2')\,d\tau_2\,d\tau_1'\,d\mathbf{p}_2'. \qquad \ldots\ldots(7.8)$$

Here we must bear in mind that the δ function of the energy refers to the true energy, i.e. that defined by (2.5). Because of this, when the function $n_0(\epsilon_0)$ is substituted in the collision integral, it does not reduce to zero. This occurs only when the true equilibrium function $n_0(\epsilon)$ is substituted in the integral. The total distribution function can be written as

$$n = n_0(\epsilon) - \frac{\partial n_0}{\partial \epsilon_0}\int f\,\delta n\,d\tau + \delta n. \qquad \ldots\ldots(7.9)$$

We introduce the notation

$$\delta n = -\frac{n_0(1-n_0)}{T}\nu = \frac{\partial n_0}{\partial \epsilon_0}\nu. \qquad \ldots\ldots(7.10)$$

Substituting in (7.9) we find that the increment to the equilibrium distribution function also has the form (7.10), but instead of ν we have

$$\psi = \nu + \left(\frac{d\tau}{d\epsilon_0}\right)_\mu \int f(\theta)\nu\frac{d\Omega}{4\pi}. \qquad \ldots\ldots(7.11)$$

Substituting (7.9) in the collision integral and neglecting quadratic terms in ψ, we find

$$I(n) = \frac{1}{T}\int w n_{01} n_{02}(1-n_{01}')(1-n_{02}')(\psi_1+\psi_2-\psi_1'-\psi_2')$$
$$\times\,\delta(\mathbf{p}_1+\mathbf{p}_2-\mathbf{p}_1'-\mathbf{p}_2')\,\delta(\epsilon_1+\epsilon_2-\epsilon_1'-\epsilon_2')\,d\tau_2\,d\tau_1'\,d\mathbf{p}_2'. \qquad \ldots\ldots(7.12)$$

The collision probability w depends in general on all four momenta, but since we are interested only in momenta with magnitudes near to that at the Fermi surface

we may assume that w depends only on θ, the angle between $\mathbf{p}_1$ and $\mathbf{p}_2$, and on ϕ, the angle between the planes formed by $(\mathbf{p}_1, \mathbf{p}_2)$ and $(\mathbf{p}_1', \mathbf{p}_2')$.

We now use the fact that the particles in the basic region of integration have momenta whose magnitudes differ only a little from the momentum p_0 at the Fermi surface. If we rotate the plane of the vectors $(\mathbf{p}_1', \mathbf{p}_2')$ through an angle ϕ about an axis along $\mathbf{p}_1+\mathbf{p}_2$, so that this plane coincides with that defined by $(\mathbf{p}_1, \mathbf{p}_2)$ we obtain the diagram shown in figure 1. It is clear that the vector $\mathbf{f}$

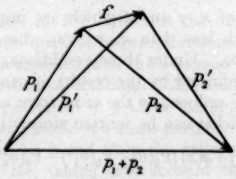


Figure 1.

will be small and that the angle between all the momenta and the axis $\mathbf{p}_1+\mathbf{p}_2$ will be approximately $\theta/2$ or $-\theta/2$. Thus we get

$$p_1' \simeq p_1 + f_s \cos\tfrac{1}{2}\theta + f_r \sin\tfrac{1}{2}\theta, \quad p_2' \simeq p_2 - f_s \cos\tfrac{1}{2}\theta + f_r \sin\tfrac{1}{2}\theta \qquad \ldots\ldots(7.13)$$

where f_s is the component of $\mathbf{f}$ along the axis $\mathbf{p}_1+\mathbf{p}_2$, and f_r is the component perpendicular to this axis. By integrating with respect to $d\mathbf{p}_2'$ we get rid of the momentum δ function, and we replace the integral over $d\tau_1'$ by one over $df_r\, df_s\, d\phi$, introducing cylindrical coordinates with axis along $\mathbf{p}_1+\mathbf{p}_2$,

$$d\tau_1' = 2p_0 \sin\tfrac{1}{2}\theta\, df_r\, df_s\, d\phi/(2\pi\hbar)^3. \qquad \ldots\ldots(7.14)$$

We now introduce the following notation

$$x = (\epsilon_1' - \mu)/T, \quad y = (\epsilon_2' - \mu)/T, \quad t = (\epsilon_1 - \mu)/T, \quad \kappa = \mu/T, \qquad \ldots\ldots(7.15)$$

with ϵ_i defined by (3.1). The change to the variables x and y instead of f_r and f_s is carried out simply by using (7.13) and we then get

$$d\tau_1' \int d\tau_2\, \delta(\epsilon) = \frac{m^{*3}\, T^2\, dx\, dy (d\Omega/2\pi)(d\phi_2/2\pi)}{8\pi^4 \hbar^6 \cos\tfrac{1}{2}\theta} \qquad \ldots\ldots(7.16)$$

where $d\Omega$ denotes the angular differential $\sin\theta\, d\theta\, d\phi$ on which $w(\theta, \phi)$ depends. We note that because the particles are indistinguishable the angle ϕ varies only from 0 to π.

From symmetry considerations it is clear that ψ must be of the form

$$\psi = \tfrac{1}{2} q(t) \left(p_i \frac{\partial \epsilon}{\partial p_x} - \frac{1}{3} p_l \frac{\partial \epsilon}{\partial p_l} \delta_{ik} \right) \left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} - \frac{2}{3} \delta_{ik} \frac{\partial u_m}{\partial x_m} \right). \qquad \ldots\ldots(7.17)$$

When this expression is substituted in the collision integral the second factor can be transformed by using the combination theorem for spherical harmonics. On integration with respect to $d\phi_2$ only the first terms remain, i.e.

$$P_2(\theta_1 + \theta_2) \to P_2(\theta_1) P_2(\theta_2) \quad \text{where} \quad P_2(\theta) = \tfrac{3}{2}\cos^2\theta - 1.$$

As a result (7.3) gives the following equation for q:

$$n_0(t)\,[1-n_0(t)] = \frac{m^{*3}\,T^2}{8\pi^4\,\hbar^6\cos\frac{1}{2}\theta}\int\frac{d\Omega}{2\pi}\int_{-\kappa}^{\infty}dx\int_{-\kappa}^{\infty}dy\,w(\theta,\phi)\,n_0(t)\,n_0(x+y-t)\,[1-n_0(x)]$$
$$\times\,[1-n_0(y)]\,[q(t)+q(x+y-t)\,\mathrm{P}_2(\theta)-q(x)\,\mathrm{P}_2(\theta_1')-q(y)\,\mathrm{P}_2(\theta_2')]. \qquad \ldots\ldots(7.18)$$
$$x+y>t-\kappa$$

Since the only values of x, y and t which are important in the temperature range considered are much less than κ, we can change the lower limit of the integration in (7.12) to $-\infty$. Under these conditions, assuming that q is a symmetric function (this is confirmed by the result), it can easily be seen that all the terms for different q can be reduced to the same form as regards their dependence on x and y. Thus the bracket can be written simply in the form

$$q(t)+q(x)\,[\mathrm{P}_2(\theta)-\mathrm{P}_2(\theta_1')-\mathrm{P}_2(\theta_2')].$$

The transform from the angles θ_1' and θ_2' to θ and ϕ is easily carried out with the help of the diagram. Integrating with respect to y, of which q is now independent, we get, after a certain amount of simplification, the following equation

$$\frac{8\pi^4\hbar^6}{m^{*3}\,T^2} = \left[\frac{w(\theta,\phi)}{\cos\frac{1}{2}\theta}\{\tfrac{3}{4}(1-\cos\theta)^2\sin^2\phi-1\}\right]_{\mathrm{av}}\left\{\int_0^{\infty}\frac{dx\,x[q(x+t)+q(x-t)]}{e^x-1}\right.$$
$$\left.+2\int_0^{\infty}\frac{dx\,x\,q(x)}{e^x+1}+\int_0^{t}dx\,x\,q(x-t)\right\}+\left[\frac{w(\theta,\phi)}{\cos\frac{1}{2}\theta}\right]_{\mathrm{av}}q(t)\,\frac{\pi^2+t^2}{2} \qquad \ldots\ldots(7.19)$$

where the suffix av denotes an average over the angles.

The equation obtained is complicated, but analysis shows that for any assumed form of $w(\theta,\phi)$ the error in the coefficient of viscosity will be less than 10%, provided we assume that the necessary values of $t^2\ll\pi^2$. If this holds, q is a constant

$$q = \frac{64\pi^2\hbar^6}{3m^{*3}\,T^2}\left\{\left[\frac{w(\theta,\phi)}{\cos\frac{1}{2}\theta}(1-\cos\theta)^2\sin^2\phi\right]_{\mathrm{av}}\right\}^{-1}. \qquad \ldots\ldots(7.20)$$

The momentum flux is given by (6.3). Substituting in this $n=n_0+\delta n$ and using (7.10) and noting that the energy is also a functional of the distribution function, we find

$$\Pi_{ik} = \int p_i\frac{\partial\epsilon(p)}{\partial p_k}\left[\delta n-\frac{\partial n_0}{\partial\epsilon_0}\int f(\mathbf{p},\mathbf{p}')\,\delta n'\,d\tau'\right]d\tau$$
$$= \int p_i\frac{\partial\epsilon(p)}{\partial p_k}\frac{\partial n_0}{\partial\epsilon_0}\left[\nu+\int f\nu'\left(\frac{d\tau'}{d\epsilon_0'}\right)_{\epsilon'=\mu}\frac{d\Omega}{4\pi}\right]d\tau = \int p_i\frac{\partial\epsilon(p)}{\partial p_k}\frac{\partial n_0}{\partial\epsilon_0}\psi\,d\tau. \qquad \ldots\ldots(7.21)$$

Thus it can be seen that the same function ψ appears in both the expression for the momentum flux and in the kinetic equation. The function f does not appear in the expression for the momentum flux and it has the same form as for a Fermi gas of particles with mass m^* and a scattering law described by the function $w(\theta,\phi)$.

Substituting (7.17) and (7.20) in (7.21) and defining the coefficient of viscosity as the coefficient of proportionality between Π_{ik} and

$$-\left(\frac{\partial u_i}{\partial x_k}+\frac{\partial u_k}{\partial x_i}-\frac{2}{3}\frac{\partial u_l}{\partial x_l}\right),$$

we find

$$\eta=\frac{64}{45}T^{-2}\frac{\hbar^3 p_0^5}{m^{*4}}\left\{\left[\frac{w(\theta,\phi)}{\cos\frac{1}{2}\theta}(1-\cos\theta)^2\sin^2\phi\right]_{av}\right\}^{-1}. \qquad \ldots\ldots(7.22)$$

The viscosity is proportional to T^{-2}, which is the dependence suggested earlier by Pomeranchuk (1950) from qualitative considerations. The numerical values of the viscosity depend on the knowledge of the averaged function $w(\theta,\phi)$ and therefore cannot be exactly determined, although (7.22) enables us to find the order of magnitude of η. As will be shown in § A2 of the appendix, the function f is of the order of the scattering amplitude of a quasi-particle. Consequently, $w\sim 2\pi f^2/\hbar$, and the magnitude of f can be found from the experimental data on the velocity of sound (see § 9). For ^{3}He, taking into account the numerical values of m^* and p_0 (4.4) we get

$$\eta=\frac{\alpha}{T^2} \quad (\alpha\sim 10^{-6}\text{–}10^{-5}\text{ poise, } T\text{ in }^\circ\text{K}). \qquad \ldots\ldots(7.23)$$

This agrees with the experimental results of Zinov'eva (1958). Since the measurements were made down to only 0·35°K, a comparison of the temperature dependence cannot be made, but the experimental results do show an increase in viscosity as the temperature is lowered.

As was noted at the beginning of this section, there is no term proportional to div **u** (with an accuracy to terms of order $(T/\mu)^2$) on the left-hand side of the kinetic equation. This means that the coefficient of second viscosity is approximately $(T/\mu)^2$ times smaller than η and is not significant in physical phenomena (for example, in the absorption of sound).

§ 8. Thermal Conductivity

We now turn to the calculation of the thermal conductivity (Abrikosov and Khalatnikov 1957 b). Assuming a small temperature gradient, we obtain for the left-hand side of the kinetic equation

$$-\frac{\partial n_0}{\partial\epsilon_0}\left(\frac{\epsilon_0-\mu}{T}-s\right)\frac{\partial\epsilon_0}{\partial\mathbf{p}}\nabla T. \qquad \ldots\ldots(8.1)$$

The transformations of the collision integral are initially the same as those of the last section (equations (7.8) to (7.16)). In this case we look for a function ψ in the form

$$\psi=q(t)\frac{\partial\epsilon_0}{\partial\mathbf{p}}\nabla T. \qquad \ldots\ldots(8.2)$$

Instead of equation (7.18) we get a similar equation with $\cos\theta$ instead of $P_s(\theta)$ in the collision integral, and with a factor $(\epsilon_0-\mu)/T-s$ on the left-hand side. The presence of this factor shows that the desired function must be made up of two terms which are symmetrical and antisymmetrical in t, i.e.

$$q(t)=q_s(t)+q_a(t). \qquad \ldots\ldots(8.3)$$

Substitution in the collision integral gives two equations

$$\frac{8\pi^4\hbar^6}{m^{*3}T^2}s = \left[\frac{w(\theta,\phi)}{\cos\frac{1}{2}\theta}\right]_{av}\left[q_a(t)\frac{\pi^2+t^2}{2} - \int_0^\infty dx\,\frac{x[q_a(x+t)+q_a(x-t)]}{e^x-1}\right.$$
$$\left. -2\int_0^\infty \frac{dx\,x\,q_a(x)}{e^x+1} - \int_0^\infty dx\,x\,q_a(x-t)\right] \qquad \ldots\ldots(8.4)$$

$$\frac{8\pi^4\hbar^6}{m^{*3}T^2} = \left[\frac{w(\theta,\phi)(1+2\cos\theta)}{\cos\frac{1}{2}\theta}\right]_{av}\left[\int_0^\infty dx\,\frac{x[q_a(x-t)-q_a(x+t)]}{e^x+1} + 2t\int_0^\infty \frac{dx\,q_a(x)}{e^x+1}\right.$$
$$\left. + \int_0^t dx\,x\,q_a(x-t)\right] + \left[\frac{w(\theta,\phi)}{\cos\frac{1}{2}\theta}\right]_{av} q_a(t)\frac{\pi^2+t^2}{2}. \qquad \ldots\ldots(8.5)$$

However, in this case the kinetic equation does not provide a complete solution of the problem. We must use an additional condition expressing conservation of the flux of mass, i.e.

$$\delta\int n\frac{\partial\epsilon}{\partial\mathbf{p}}d\tau = \int\delta n\frac{\partial\epsilon_0}{\partial\mathbf{p}}d\tau + \int n_0\frac{\partial}{\partial\mathbf{p}}\left[\int f\,\delta n'\,d\tau'\right]d\tau$$
$$= \int\frac{\partial\epsilon_0}{\partial\mathbf{p}}\left[\delta n - \frac{\partial n_0}{\partial\epsilon_0}\int f\,\delta n'\,d\tau'\right]d\tau = 0. \qquad \ldots\ldots(8.6)$$

Here we have integrated by parts and replaced ϵ by ϵ_0 in the integrals containing δn. Substituting (7.10), (7.11) and (8.2) in (8.6), we find

$$\int\frac{\partial n_0}{\partial\epsilon_0}\left(\frac{\partial\epsilon_0}{\partial\mathbf{p}}\right)^2 q(t)\,d\tau = 0. \qquad \ldots\ldots(8.7)$$

To find $q_a(t)$ it is sufficient to solve (8.5) in a way similar to that used in the discussion of the viscosity, but the situation is quite different as regards the even part $q_a(t)$. First, we note that $q_a = \text{const.}$ makes the right-hand side of (8.4) identically zero, and therefore the constant term $q_a(0)$ must be determined from the condition (8.7) and not from this equation. It is also easy to show that the further terms in $q_a(t)$, i.e. $a_2t^2 + a_4t^4 \ldots$ do not in general contribute to the thermal conductivity. The appearance of a term such as, for example, $a_m t^{2m}$, must change the constant term in $q_a(t)$ to $a_m{}^0$, such that

$$\int\frac{\partial n_0}{\partial\epsilon_0}\left(\frac{\partial\epsilon}{\partial\mathbf{p}}\right)^2[a_m t^{2m} + a_m{}^0]\,d\tau = 0.$$

From this condition we get

$$a_m{}^0 = -a_m(2m)!\,R_m \qquad R_m = \int_0^\infty \frac{z^{2m-1}}{e^z+1}dz.$$

To find the energy flux we must calculate integrals of the type

$$\int\frac{\partial n_0}{\partial\epsilon_0}F(\epsilon_0)(a_m t^{2m} + a_m{}^0)\,d\tau = \left(F\frac{d\tau}{d\epsilon_0}\right)_{\epsilon_0=\mu}[a_m(2m)!\,R_m + a_m{}^0] = 0.$$

Thus we arrive at the conclusion that to find the coefficient of thermal conductivity we need solve only (8.5), choosing the constant term to satisfy the condition (8.7).

In practice, as for the viscosity case, it is sufficient to find a solution assuming $t^2 \ll \pi^2$. Then we get

$$q = \frac{24\pi^2\hbar^6}{m^{*3}T^2}\left(t - \frac{\pi^2 Tm^*}{{p_0}^2}\right)\left\{\left[\frac{w(\theta,\phi)(1-\cos\theta)}{\cos\tfrac{1}{2}\theta}\right]_{\mathrm{av}}\right\}^{-1}. \qquad \ldots\ldots(8.8)$$

The energy flux is given by (6.5). Substituting (7.1) in this equation, we have

$$Q = \int \epsilon_0(\mathbf{p})\frac{\partial\epsilon_0(\mathbf{p})}{\partial\mathbf{p}}\,\delta n\,d\tau - \int \epsilon_0(\mathbf{p})\frac{\partial\epsilon_0(\mathbf{p})}{\partial\mathbf{p}}\frac{\partial n_0}{\partial\epsilon_0}\int f\,\delta n'\,d\tau'\,d\tau. \qquad \ldots\ldots(8.9)$$

Using (7.10), (7.11), (8.2) and (8.8), we find the coefficient of thermal conductivity

$$\kappa = \frac{8}{3}\frac{\pi^2\hbar^3 {p_0}^3}{m^{*4}T}\left\{\left[\frac{w(\theta,\phi)(1-\cos\theta)}{\cos\tfrac{1}{2}\theta}\right]_{\mathrm{av}}\right\}^{-1}. \qquad \ldots\ldots(8.10)$$

As for the viscosity, the coefficient κ does not contain f. The thermal conductivity is inversely proportional to the temperature, which also agrees with the qualitative predictions of Pomeranchuk (1950).

Estimating the numerical value of the thermal conductivity of ^{3}He by a method analogous to that used for the viscosity, we obtain the relation

$$\kappa = \beta/T \quad (\beta \sim 10^2\text{–}10^3\ \mathrm{erg\,cm^{-1}\,sec^{-1}\,deg^{-1}},\ T \text{ in } {}^\circ\mathrm{K}). \qquad \ldots\ldots(8.11)$$

The expressions obtained for the viscosity and thermal conductivity cease to be valid for $T \sim \mu$. However, there is another condition which limits the range of application of the theory to much lower temperatures. This is that the excitation energies, which are of the order T, must be considerably greater than the quantum indeterminacy in the energy which is due to collisions, i.e.

$$\tau \gg \hbar/T \qquad \ldots\ldots(8.12)$$

where τ is the time between collisions. We note that the condition (8.12) must be fulfilled not just for the calculation of the kinetic coefficients, but also for the whole theory of a Fermi liquid to be valid.

For ^{3}He the magnitude of τ can be determined from the experimental values of the viscosity (Zinov'eva 1958) (see formula (10.12)). If we use the number at the lowest temperature available, i.e. $\eta = 4{\cdot}8 \times 10^{-5}$ poise at $T = 0{\cdot}35^\circ\mathrm{K}$, then from (8.12) we get

$$T \ll 0{\cdot}3^\circ\mathrm{K}. \qquad \ldots\ldots(8.13)$$

However, it is clear that this value is rather high, as for $T = 0{\cdot}35^\circ\mathrm{K}$ the viscosity is still not proportional to T^{-2} and varies more slowly than this.

§9. Sound

The propagation of sound in a Fermi liquid has several quite specific peculiarities. If we consider sound of a given frequency, then provided the temperature is not too low it will propagate according to the laws of normal hydrodynamics and its damping will be proportional to τ. As the temperature drops the collision time will increase as T^{-2}, and finally, when $\omega\tau \sim 1$, sound will be completely damped. However, if the temperature is further reduced sound propagation becomes possible again. Its velocity will be different and it will not simply be a wave

of compression and rarefaction. This phenomenon was predicted by Landau (1957) and was named by him 'zero sound'. Since only the relation between ω and τ is important in the determination of the kind of sound, these two sounds can be described as low-frequency sound ($\omega\tau \ll 1$), and high-frequency sound ($\omega\tau \gg 1$).

The velocity of sound at low frequencies or at temperatures which are not too low, when the condition $\omega\tau \ll 1$ is fulfilled, is given in the usual way in terms of the compressibility. Landau (1956) showed that it depends considerably on the function f.

It is convenient to express the compressibility in terms of the derivative $\partial\mu/\partial N$. Here N is the total number of particles. Since the chemical potential μ depends only on N/V, we have

$$\frac{\partial\mu}{\partial N} = -\frac{V\partial\mu/\partial V}{N} = -\frac{V^2}{N^2}\frac{\partial p}{\partial V}. \qquad \ldots\ldots(9.1)$$

The velocity of sound is determined by the relation

$$c^2 = \frac{\partial p}{\partial \rho} = \frac{\partial p}{\partial (mN/V)} = \frac{1}{m}\left(N\frac{\partial\mu}{\partial N}\right). \qquad \ldots\ldots(9.2)$$

The derivative $\partial\mu/\partial N$ is calculated in the following way. Since $\mu = \epsilon(p_0)$, a change in μ results both from a change in p_0 and from a change in the form of the function $\epsilon(p)$

$$\delta\mu = \int f\,\delta n'\,d\tau' + \frac{\partial\epsilon_0}{\partial p_0}\delta p_0. \qquad \ldots\ldots(9.3)$$

δN and δp are related by the equation

$$\delta N = 8\pi p_0^2\,\delta p_0\,V/(2\pi\hbar)^3. \qquad \ldots\ldots(9.4)$$

Since only the changes δn near the Fermi surface are important in the integral in (9.3), we may carry out the integration with respect to the magnitude of the momentum and this gives

$$\int f\,\delta n'\,d\tau' = \frac{\delta N}{4\pi V}\int f\,d\Omega. \qquad \ldots\ldots(9.5)$$

Substitution in (9.3) gives (see also (3.1) and (3.2))

$$\frac{\partial\mu}{\partial N} = \frac{1}{4\pi V}\int f\,d\Omega + \frac{(2\pi\hbar)^3}{8\pi p_0 m^* V}. \qquad \ldots\ldots(9.6)$$

(3.5) enables us to express the effective mass m^* in terms of the true mass of the atoms m. Substituting the expression obtained into (9.2), and using

$$N = 8\pi p_0^3\,V/3(2\pi\hbar)^3,$$

we find

$$c^2 = \frac{p_0^2}{3m^2} + \frac{2}{3m}\left(\frac{p_0}{2\pi\hbar}\right)^3\int f(\chi)(1-\cos\chi)\,d\Omega. \qquad \ldots\ldots(9.7)$$

(3.5) and (9.7) give us some information about the function $f(\chi)$. If we introduce the notation

$$F(\chi) = f(\chi)\left(\frac{d\tau}{d\epsilon}\right)_{\epsilon=\mu} = f(\chi)\frac{p_0 m^*}{\pi^2\hbar^3} = \sum_n F_n P_n(\cos\chi), \qquad \ldots\ldots(9.8)$$

the first coefficients of the expansion in Legendre polynomials can be found from (3.5) and (9.7). It is not difficult to obtain the relations

$$\frac{1}{m} = \frac{1}{m^*}(1 + [F\cos\chi]_{\mathrm{av}}),$$

$$c^2 = \frac{p_0^2}{3m^2}\frac{1+[F]_{\mathrm{av}}}{1+[F\cos\chi]_{\mathrm{av}}}. \qquad \ldots\ldots(9.9)$$

Using the values of the parameters for ^{3}He from (4.4) and the value of the velocity of sound (Lagner, Sydoriak and Roberts 1957), $c_{T=0} = 183\ \mathrm{m\,sec^{-1}}$, we find

$$F_0 = [F]_{\mathrm{av}} = 4{\cdot}65, \quad F_1 = 3[F\cos\chi]_{\mathrm{av}} = 1{\cdot}3. \qquad \ldots\ldots(9.10)$$

Since the remaining coefficients cannot be determined so simply we shall in further calculations use only the first two harmonics of the function F

$$F = F_0 + F_1\cos\chi. \qquad \ldots\ldots(9.11)$$

We shall now investigate 'zero sound', i.e. the case $\omega\tau \gg 1$. The kinetic equation takes the form (6.1) and in it we substitute the distribution function in the form

$$n = n_0 + \delta n \qquad \ldots\ldots(9.12)$$

where the small deviation δn is periodic in both space and time: $\delta n \sim \exp[i(\mathbf{kr} - \omega t)]$. According to (2.5) the energy of the excitations also changes and this must be taken into account. Since the collision integral $I(n) \sim \delta n/\tau$, and $\omega\tau \gg 1$, we may neglect $I(n)$ and hence we have

$$(\mathbf{kv} - \omega)\,\delta n - \mathbf{kv}\frac{\partial n_0}{\partial \epsilon}\int f\,\delta n'\,d\tau' = 0.$$

From the form of this equation it follows that δn is proportional to $\partial n_0/\partial\epsilon$; putting $\delta n = \nu\partial n_0/\partial\epsilon$, we find

$$(\mathbf{kv} - \omega)\nu + \mathbf{kv}\int F\nu'\frac{d\Omega}{4\pi} = 0. \qquad \ldots\ldots(9.13)$$

If we choose $\mathbf{k}$ as a polar axis and introduce the notation $u = \omega/k$, the velocity of propagation of the waves and $s = u/v$, then equation (9.13) becomes

$$(s - \cos\theta)\nu(\theta,\phi) = \cos\theta\int F(\chi)\nu(\theta',\phi')\,d\Omega/4\pi. \qquad \ldots\ldots(9.14)$$

Hence we can find $\nu(\theta,\phi)$ and the value of s. Since the change in the distribution function δn is proportional to $\partial n_0/\partial\epsilon$ it is clear that the phenomenon reduces to a distortion of the Fermi surface, the form of which depends on the function ν. Equation (9.14) depends considerably on the function F.

First, as an example, we consider the simplest possible case, $F = F_0 = \mathrm{const.}$ Then (9.14) gives (dropping the periodic exponential factor)

$$\nu = \frac{\mathrm{const.}\cos\theta}{s - \cos\theta}. \qquad \ldots\ldots(9.15)$$

Displacement of the Fermi surface as a whole (this corresponds to ordinary sound) would correspond to $\nu \sim \cos\theta$, thus here we are dealing with an anisotropic distortion of the Fermi surface. As we shall see later, s must be greater than 1 and thus the Fermi surface is extended in the direction of the motion.

Substituting (9.15) into (9.14) with $F = F_0$ we find an equation for s, and after an elementary integration we have

$$\tfrac{1}{2}s \ln\left(\frac{s+1}{s-1}\right) - 1 = \frac{1}{F_0}. \qquad \ldots\ldots(9.16)$$

Hence it is clear that if s is real (corresponding to undamped waves) then it must be greater than 1, i.e.

$$u > v. \qquad \ldots\ldots(9.17)$$

(From (9.14) it is clear that this condition remains valid for any function F). The left-hand side of (9.16) being always positive, the condition for the existence of zero sound is evidently $F_0 > 0$.

If the function F_0 is large, then so is s, and from (9.16) we get $s \to (F_0/3)^{1/2}$ as $F_0 \to \infty$. On the other hand, as $F_0 \to 0$ $s \to 1$, i.e. $u \to v$; this is the case of an almost free Fermi gas.

It is not difficult to see that the conclusion $s \to 1$ as $F \to 0$ is independent of the form of F. From (9.14) it follows indeed that $s \to 1$ for $F \to 0$, and that ν is non-zero only for small θ. From (9.7) we find that for a Fermi gas which is nearly ideal, $c^2 \simeq p_0^2/3m$, i.e. $c \simeq v/\sqrt{3} \simeq u/\sqrt{3}$. Thus the velocity of zero sound will exceed that of ordinary sound by a factor of $\sqrt{3}$.

We now consider the general case of $F(\chi)$ (Abrikosov and Khalatnikov 1957 c). For that purpose we write this function as a sum of a series of spherical harmonics, according to (9.8). Substituting into (9.14), and using the combination theorem for Legendre polynomials

$$\mathrm{P}_n(\chi) = \sum_{m=-n}^{n} \mathrm{P}_n^m(\theta)\,\mathrm{P}_n^m(\theta')\exp\{im(\phi-\phi')\}\frac{(n-|m|)!}{(n+|m|)!}$$

where $\mathrm{P}_n^m = \mathrm{P}_n^{-m}$ are associated Legendre polynomials, we finally get

$$(\cos\theta - s)\nu + \cos\theta \sum \frac{(n-|m|)!}{(n+|m|)!}\,\mathrm{P}_n^m(\theta)\,F_n \exp(im\phi)$$
$$\times \int \mathrm{P}_n^m(\theta')\,\nu(\theta',\phi')\exp(-im\phi')\frac{d\Omega'}{4\pi} = 0. \qquad \ldots\ldots(9.18)$$

We introduce the notation

$$F_n \frac{(n-|m|)!}{(n+|m|)!}\int \mathrm{P}_n^m(\theta')\,\nu(\theta',\phi')\exp(-im\phi')\frac{d\Omega}{4\pi} = \Phi_{nm} \qquad \ldots\ldots(9.19)$$

and solve (9.18) for

$$\nu = -\frac{\cos\theta}{\cos\theta - s}\sum \Phi_{nm}\,\mathrm{P}_n^m(\theta)\,\mathrm{e}^{im\phi}. \qquad \ldots\ldots(9.20)$$

Substituting this expression in (9.19) and integrating with respect to ϕ', we get

$$F_n \frac{(n-|m|)!}{(n+|m|)!}\int \sum_k \mathrm{P}_n^m(\theta')\frac{\cos\theta'}{\cos\theta' - s}\,\mathrm{P}_k^m(\theta')\frac{d\Omega}{4\pi}\,\Phi_{km} = \sum_k \Phi_{km}\,\delta_{kn}. \qquad \ldots\ldots(9.21)$$

Thus we have obtained a system of homogeneous equations which determine the quantities Φ_{km}. The system separates into independent sub-systems corresponding to various values of m. Equation (9.21) shows that oscillations of several different types can propagate in a Fermi liquid at absolute zero and that they are characterized by a different dependence of the amplitude on the angles θ, ϕ. The value $m = 0$ corresponds to oscillations for which ν is isotropic in a plane perpendicular to $\mathbf{k}$, while for $m \neq 0$ the oscillations are polarized in a definite way in this plane. The number of types of oscillations is determined by the number of possible values of m ($|m| \leqslant n$). The velocity of propagation of the oscillations is found from the condition that the determinant of the corresponding system should be zero:

$$\| \delta_{kn} + F_n \Omega^m_{kn}(s) \| = 0 \quad (N \geqslant n, k \geqslant m)$$

$$\Omega^m_{kn}(s) = \frac{(n-|m|)!}{(n+|m|)!} \int \mathrm{P}^m_k(\theta') \frac{\cos\theta'}{\cos\theta' - s} \mathrm{P}^m_n(\theta') \frac{d\Omega}{4\pi}. \qquad \ldots\ldots(9.22)$$

Since $\mathrm{P}^{-m}_n = \mathrm{P}^m_n$, the coefficients Ω^m_{kn} do not depend on the sign of m, and therefore oscillations which differ only in the sign of m propagate with the same velocity.

Equation (9.22) shows that the equations for the velocity are transcendental. In the general case they do not always have real roots, but cases are possible in which there are several. This corresponds to several types of oscillation with identical polarization in a plane perpendicular to $\mathbf{k}$.

As an example we consider the case when the function $F(\chi)$ contains only the zeroth and first harmonic (function (9.11)). The coefficients Ω^m_{kn} are given by

$$\Omega^0_{00} = \frac{1}{2}\int_{-1}^{1} \frac{x\,dx}{x-s} = 1 - \frac{s}{2}\ln\left(\frac{s+1}{s-1}\right) = -w,$$

$$\Omega^0_{10} = \Omega^0_{01} = \frac{1}{2}\int_{-1}^{1} \frac{x^2\,dx}{x-s} = -sw,$$

$$\Omega^0_{11} = \frac{1}{2}\int_{-1}^{1} \frac{x^3\,dx}{x-s} = \frac{1}{3} - s^2 w,$$

$$\Omega^1_{11} = \frac{1}{4}\int_{-1}^{1} \frac{(1-x^2)x\,dx}{x-s} = \tfrac{1}{2}[(s^2-1)w - \tfrac{1}{3}].$$

On substitution in the determinant (9.22), we get the following equation for the velocity of propagation of the oscillations of the type $m = 0$:

$$w = \frac{1 + \frac{1}{3}F_1}{F_0 + \frac{1}{3}F_0F_1 + F_1 s^2}. \qquad \ldots\ldots(9.23)$$

For $m = 1$ we get the equation

$$w = \frac{F_1 - 6}{3F_1(s^2-1)} \qquad \ldots\ldots(9.24)$$

which has one real root for $F_1 > 6$.

We now discuss the application of the formulae obtained to liquid helium 3. From equation (9.10) it follows that all the oscillations of type $m = 0$ have the same velocity (the root of (9.23)): $s = u/v = 1{\cdot}72$, $u = 192\ \mathrm{m\,sec^{-1}}$. There are

no oscillations of the type $m = 1$ (or for $m > 1$). It is, of course, possible that this deduction is a consequence of the roughness of the approximation which was used for $F(\chi)$, but we see no basis for this conclusion.

As well as the oscillations considered, waves of another kind, which can be called spin waves, may propagate in a Fermi liquid at absolute zero (Landau 1957). In the discussion of such oscillations it is necessary to take into account the spin-dependent part of the function f, as determined by (5.3). We introduce the new notation

$$\Phi(\chi) = f(\chi)\left(\frac{d\tau}{d\epsilon}\right)_{\epsilon=\mu} = F(\chi) + Z(\chi)\,\boldsymbol{\sigma}\boldsymbol{\sigma}' \qquad \ldots\ldots(9.25)$$

and the equation for zero sound now becomes

$$(s - \cos\theta)\,\nu = \cos\theta \cdot \tfrac{1}{2}\mathrm{Sp}_{\sigma'} \int \Phi\nu'\, d\Omega'/4\pi. \qquad \ldots\ldots(9.26)$$

If we take ν as independent of the spin, this goes over into (9.14); but ν can also be of another form, i.e.

$$\nu = \boldsymbol{\nu}\boldsymbol{\sigma}, \qquad \ldots\ldots(9.27)$$

and for this function we get the equation

$$(s - \cos\theta)\,\boldsymbol{\nu} = \cos\theta \int Z\boldsymbol{\nu}'\, d\Omega'/16\pi. \qquad \ldots\ldots(9.28)$$

Thus for the component of the vector $\boldsymbol{\nu}$ we get an equation which differs from (9.14) only by the replacement of F by $Z/4$. Thus all further calculations must be the same as those already made in § 9.

As was shown in § 5, Z_0 for helium 3 is approximately $-3{\cdot}4$, i.e. negative. Although it is not possible to derive any rigorous conclusions from this it is most probable that spin waves cannot propagate in ^{3}He.

The possibility of the propagation of sound waves at $T = 0$ means that a Fermi liquid has a Bose branch of the energy spectrum in the form of phonons with energy $\epsilon = up$. However, the corrections to the thermodynamical quantities due to this branch contain higher powers of T (specific heat $\sim T^3$) which are neglected in the present approximation of the theory.

§ 10. Dispersion and Absorption of Sound

In the previous section the two limiting cases of $\omega\tau \ll 1$ and $\omega\tau \gg 1$ were considered. We now consider the transition between ordinary and zero sound, i.e. dispersion region (Abrikosov and Khalatnikov 1957 c). Since the region of dispersion is at the same time that of strong absorption, our discussion enables us to solve simultaneously the problem of absorption of sound.

We must solve the kinetic equation, including the collision integral, but since the law of interaction of the excitations is unknown, we shall simplify the form of the collision integral.

To do this we could introduce an effective time τ and replace the collision integral in (6.1) by the expression $-\delta n/\tau$. However, if this substitution were made, the conservation laws for the number of excitations, momentum and energy would not follow, and this would make the transition to hydrodynamics impossible.

Since both the second viscosity and the effect of thermal conductivity (see below) can be neglected here, the collision integral does not contain terms in the zeroth and first spherical harmonics. We therefore replace the collision integral by the following expression†

$$I(n) \to -\tau^{-1}(\delta n - [\delta n]_{av} - 3[\delta n \cos\theta]_{av} \cos\theta). \qquad \ldots\ldots(10.1)$$

It is easy to see that on integration with respect to $d\tau$ this expression reduces to zero. It also reduces to zero on multiplication by ϵ or by $p\cos\theta$ and integration with respect to $d\tau$.‡ Thus the equations for the conservation of energy, momentum and the total number of particles will automatically be satisfied. Thus, using (9.14) and (10.1), the initial kinetic equation becomes

$$(kv\cos\theta - \omega)\nu + kv\cos\theta \int F\nu' \frac{d\Omega}{4\pi} = -\frac{1}{i\tau}(\nu - [\nu]_{av} - 3[\nu\cos\theta]_{av}\cos\theta). \qquad \ldots\ldots(10.2)$$

Bearing in mind the application of the results to ^{3}He, and to simplify the problem, we write the function F in the form (9.11).

We introduce the notation

$$[\nu]_{av} = \nu_0, \quad 3[\nu\cos\theta]_{av} = \nu_1, \quad \sigma = i\tau kv, \quad \xi = (i\omega\tau - 1)/i\tau kv, \qquad \ldots\ldots(10.3)$$

after which (10.2) easily reduces to

$$(\cos\theta - \xi)\nu + \cos\theta(F_0\nu_0 + \tfrac{1}{3}F_1\nu_1\cos\theta) = \frac{1}{\sigma}(\nu_0 + \nu_1\cos\theta). \qquad \ldots\ldots(10.4)$$

We now solve this equation for ν and calculate $[\nu]_{av} = \nu_0$ and $[\nu\cos\theta]_{av} = \nu_1/3$. Thus we get two equations for the two quantities ν_0 and ν_1,

$$\nu_0 = F_0\nu_0 w + \tfrac{1}{3}wF_1\nu_1 - \frac{1}{\sigma}\frac{w+1}{\xi}\nu_0 - \frac{\nu_1}{\sigma}w, \qquad \ldots\ldots(10.5)$$

$$\tfrac{1}{3}\nu_1 = F_0\nu_0\xi w - \tfrac{1}{3}(\tfrac{1}{3} - \xi^2 w)F_1\nu_1 - \frac{w}{\sigma}\nu_0 - \frac{\xi w}{\sigma}\nu_1, \qquad \ldots\ldots(10.6)$$

where

$$w = \tfrac{1}{2}\xi\ln\frac{\xi+1}{\xi-1} - 1.$$

From the condition that the equations (10.5) and (10.6) should be consistent, we get an equation which determines the complex velocity of sound

$$\left(1+\frac{1}{\xi\sigma}\right)\left(1+\frac{F_1}{3}\right) - w\left\{\left(1+\frac{F_1}{3}\right)\left(F_0 - \frac{1}{\xi\sigma}\right) + \xi^2\left(F_1 - \frac{3}{\xi\sigma}\right)\left(1+\frac{1}{\xi\sigma}\right)\right\} = 0. \qquad \ldots\ldots(10.7)$$

This is the desired equation which gives the frequency dependence of the velocity of sound, or, in other words, which describes the dispersion of sound in a Fermi liquid. First, we consider the two limiting cases.

† Henceforth $[\,]_{av}$ denotes an average with respect to the angle.

‡ In this integration only the values of momenta near to p_0 are important because δn contains a δ-function for $\epsilon = \mu$.

(*a*) Low frequencies: $\omega\tau \ll 1$. It follows that $\sigma \to 0$, $\xi\sigma \to -1$, $\xi \to \infty$. The expansion of w in powers of $1/\xi$ gives

$$w = \frac{1}{3\xi^2} + \frac{1}{5\xi^4},$$

and on simplification equation (10.7) becomes

$$\left(1+\frac{1}{\xi\sigma}\right)^2 = \frac{1}{3\xi^2}\left(1+\frac{F_1}{3}\right)\left(F_0 - \frac{1}{\xi\sigma}\right) + \frac{1}{5\xi^2}\left(1+\frac{1}{\xi\sigma}\right)\left(F_1 - \frac{3}{\xi\sigma}\right). \quad \ldots\ldots(10.8)$$

Taking into account (10.3), we have

$$\left(\frac{1+\xi\sigma}{\sigma}\right)^2 = \left(\frac{\omega}{kv}\right)^2.$$

From (10.8) and (10.3) we find to first order in $i\omega\tau$

$$\left(\frac{\omega}{kv}\right)^2 = \tfrac{1}{3}(1+F_0)(1+\tfrac{1}{3}F_1) - \tfrac{4}{15}i\,\omega\tau(1+\tfrac{1}{3}F_1). \quad \ldots\ldots(10.9)$$

The first term corresponds to the velocity of ordinary sound in a Fermi liquid. The absorption of sound in the region $\omega\tau \ll 1$ can easily be found from (10.9) as the imaginary part of the wave vector

$$\gamma = \mathcal{I}k = \left(\frac{2\omega^2\tau v^2}{15c^3}\right)\left(1+\frac{F_1}{3}\right). \quad \ldots\ldots(10.10)$$

Since the present case corresponds to ordinary sound, we can use the more accurate usual formula instead of (10.10)

$$\gamma = \frac{\omega^2}{2\rho c^3}\left|\left(\frac{4}{3}\eta + \zeta\right) + \frac{\kappa}{c_p}\left(\frac{c_p}{c_v} - 1\right)\right|.$$

Taking into account that $\zeta \ll \eta$ and that $\kappa/_p c$, which is of order η, is multiplied by the small factor $c_p/c_v - 1$ (it is not difficult to see that it is of order $(T/\mu)^2$), we find

$$\gamma = \frac{2\omega^2}{3\rho c^3}\eta. \quad \ldots\ldots(10.11)$$

By comparing this with (10.10) we can find a relation which enables us to find the time τ from the experimental values of the viscosity

$$\eta = \tfrac{1}{5}\rho\tau v^2(1+\tfrac{1}{3}F_1). \quad \ldots\ldots(10.12)$$

From the viscosity data of Zinov'eva (1958)

$$\tau = 2{\cdot}3 \times 10^{-12}\, T^{-2}\,\text{sec} \quad (T \text{ in } {}^\circ\text{K}) \quad \ldots\ldots(10.13)$$

and thus we get for the absorption coefficient

$$\gamma \sim 1 \times 10^{-17}(\omega/T)^2\,\text{cm}^{-1}. \quad \ldots\ldots(10.14)$$

(*b*) We now consider the second limiting case of high frequencies and low temperatures $\omega\tau \gg 1$. In this case

$$\sigma \to \infty, \quad \xi\sigma \to \infty, \quad \xi = s + i\xi', \quad |\xi'| \ll s.$$

Equation (10.7) in zeroth approximation becomes

$$(1+\tfrac{1}{3}F_1)-w(s)\{(1+\tfrac{1}{3}F_1)F_0+s^2F_1\}=0. \qquad \ldots\ldots(10.15)$$

This equation coincides with (9.23) which determined the velocity of zero sound.

To calculate the absorption of zero sound it is necessary to find the imaginary part of the velocity of sound ξ'. From (10.9) we get the equation

$$\xi'\left\{\frac{1}{w(s)}\left(1+\frac{F_1}{3}\right)\left(\frac{s}{s^2-1}+\frac{w(s)+1}{s}\right)-2sw(s)F_1\right\}$$
$$-\frac{1}{\omega\tau}\left\{\left(1+\frac{F_1}{3}\right)[1+w(s)]+w(s)s^2(3-F_1)\right\}=0. \qquad \ldots\ldots(10.16)$$

Using (10.3), we find the absorption coefficient

$$\gamma=\mathscr{I}k=\frac{1}{s\tau v}-\frac{\omega\xi'}{s^2v}. \qquad \ldots\ldots(10.17)$$

If we substitute the values of the parameters for ^{3}He ($s=1{\cdot}72, v=1{\cdot}11\times10^4\,\mathrm{cm\,sec^{-1}}$) we get

$$\xi'=0{\cdot}16/\omega\tau, \quad \gamma\sim2\times10^7\,T^2\ \mathrm{cm^{-1}}. \qquad \ldots\ldots(10.18)$$

Thus the absorption of zero sound is independent of the frequency and increases with temperature as $1/\tau$, i.e. proportional to T^2.

In these calculations sound has been considered classically and this approach is valid provided $\hbar\omega\ll T$. For $\hbar\omega\gtrsim T$ a quantum approach must be used. The reduction in the number of sound quanta per unit time due to collisions between Fermi particles is

$$\int w\{n_1n_2(1-n_1')(1-n_2')-n_1'n_2'(1-n_1)(1-n_2)\}\,\delta(\mathbf{p}_1'+\mathbf{p}_2'-\mathbf{p}_1-\mathbf{p}_2-\hbar\mathbf{k})$$
$$\times\,\delta(\epsilon_1'+\epsilon_2'-\epsilon_1-\epsilon_2-\hbar\omega)\,d\tau_1\,d\tau_2\,d\tau_1'\,d\mathbf{p}_2'. \qquad \ldots\ldots(10.19)$$

The function w of the four momenta is unknown, but it is possible to express the absorption coefficient in terms of its classical value (10.17) (this was done by Landau (1957)).

Using the fact that in the region of the smearing out of the Fermi distribution the only quickly varying functions are the occupation numbers $n(\epsilon)$, we can assume that the angular integrals give a constant which is unchanged by the transition from the classical region $\hbar\omega\ll T$ to the quantum region $\hbar\omega\gg T$. As a result we are left with only the integral over the energies

$$J=\int\{n_1n_2(1-n_1')(1-n_2')-n_1'n_2'(1-n_1)(1-n_2)\}$$
$$\times\,\delta(\epsilon_1'+\epsilon_2'-\epsilon_1-\epsilon_2-\hbar\omega)\,d\epsilon_1\,d\epsilon_2\,d\epsilon_1'\,d\epsilon_2'. \qquad \ldots\ldots(10.20)$$

Substituting the Fermi function for n and extending the integration from $-\infty$ to ∞ (this is permissible as the integrals are strongly convergent), we obtain, after a rather long calculation

$$J=T^3\frac{2\pi^2}{3}\hbar\omega\left[1+\left(\frac{\hbar\omega}{2\pi T}\right)^2\right]. \qquad \ldots\ldots(10.21)$$

The absorption coefficient is proportional to the integral J and the coefficient of proportionality can be found from the limiting classical case. Since we must obtain the classical value for $\hbar\omega \ll T$, we have in the general case

$$\gamma = \gamma_{\text{cl}}\left[1+\left(\frac{\hbar\omega}{2\pi T}\right)^2\right], \qquad \ldots\ldots(10.22)$$

where γ_{cl} is given by (10.17). Since γ_{cl} is proportional to T^2, γ must be proportional to ω^2 in the limiting quantum case $\hbar\omega \gg T$.

§ 11. Light Scattering (fluctuations in the distribution function)

It has been shown above that the condition for the possibility of propagation of 'zero sound' is the inequality

$$\omega_s \tau \gg 1 \qquad \ldots\ldots(11.1)$$

where τ is the collision time, which for ^{3}He is of the order of $10^{-12}\,T^{-2}$ sec. Even if we take a temperature of the order of 0·01°, we need a frequency of more than 10^8 c/s to observe zero sound, which makes such an experiment very difficult.

Instead of this an indirect method, which consists of an observation of Rayleigh scattering of light in liquid ^{3}He,† could be used. In Rayleigh scattering satellite lines are observed in addition to the main line, with a frequency difference

$$\Delta\omega = \pm 2\frac{u}{c}\omega\sin\tfrac{1}{2}\theta$$

where u is the velocity of sound, θ is the angle of scattering. The velocity of zero sound in ^{3}He is of the order of 2×10^4 cm sec^{-1}, i.e. $\Delta\omega \sim 10^{-6}\omega$. Thus observation of the frequency distribution of scattered light gives us a chance of measuring the velocity of zero sound. The condition (11.1) can be satisfied by the very high frequency of visible light.

Besides this aspect of the problem, the scattering of light in a Fermi liquid at a sufficiently low temperature has a series of specific features, in view of which it is interesting to carry out a theoretical investigation of this phenomenon, especially as regards the frequency distribution of its intensity (Abrikosov and Khalatnikov 1958).‡

The frequency and angular dependence for Rayleigh scattering of unpolarized light are given by the formula (see Landau and Lifshitz 1957 a)§

$$d\lambda = \frac{\omega^4}{6\pi c^4}\frac{1}{2\pi V}\left[\left|\int \delta D_{\Delta\omega}(\mathbf{r})\exp(-i\mathbf{qr})\,dV\right|^2\right]_{\text{av}} \tfrac{3}{4}(1+\cos^2\theta)\frac{d\Omega}{4\pi}\,d\Delta\omega \qquad \ldots\ldots(11.2)$$

where ω is the frequency of the incident light, θ is the angle of scattering, $\mathbf{q}$ is the change in the wave vector of the light, whose magnitude is $(2\omega/c)\sin\frac{1}{2}\theta$,

† The idea of using Rayleigh scattering was first suggested by S. P. Kapitza.

‡ We note that for high temperatures, when $\omega_s\tau \ll 1$, light scattering will be described by the usual formulae (see Landau and Lifshitz 1957 a).

§ Here $d\lambda$ is the so-called differential extinction coefficient. The integral of $d\lambda$ with respect to $d\Omega$ and $d\Delta\omega$ give the total extinction coefficient λ, which is the damping decrement of the flux density of phonons in the medium.

$\delta D_{\Delta\omega}$ is the Fourier component of the fluctuation of the dielectric constant $\delta D(t)$

$$\delta D_{\Delta\omega} = \frac{1}{\sqrt{t_0}} \int_0^{t_0} \delta D(t) \exp(i\Delta\omega t)\, dt \qquad \ldots\ldots(11.3)$$

where t_0 is some large number which will later be allowed to tend to infinity. In (11.2), $[]_{\mathrm{av}}$ denotes an average over the fluctuations. For simplicity we shall in future assume that the system occupies unit volume.

Because of the very small polarizability of helium atoms it can be assumed that the change in dielectric constant is a result of fluctuations in density, i.e. $\delta D = (\partial D/\partial N)\,\delta N$ where N is the number of particles per unit volume. But, by the general theory of a Fermi liquid, the number of excitations equals the number of atoms in the liquid, and thus we may write

$$\int \delta D_{\Delta\omega}(\mathbf{r}) \exp(-i\mathbf{qr})\, dV = \frac{\partial D}{\partial N} \int \delta n_{\mathbf{q},\Delta\omega}(\mathbf{p})\, d\tau_p \qquad \ldots\ldots(11.4)$$

where $\delta n_{\mathbf{q},\Delta\omega}(\mathbf{p})$ is the Fourier component with respect to $\mathbf{r}$ and t (the latter in the sense of equation (11.3)) of the fluctuation of the excitation distribution function.

Before embarking on further calculations, we note one important point. Equation (11.2) contains an average over all possible fluctuations, but in the temperature and frequency region where $\hbar\Delta\omega \gtrsim T$ this average must be made with due account of quantum effects. However (see Landau and Lifshitz 1957 a), it is sufficient to know the result for the purely classical case, i.e. for $T \gg \hbar\Delta\omega$, and then to introduce a correcting factor. For scattering with a frequency increase of $\Delta\omega$ (anti-Stokes scattering) we must use a factor $(\hbar\Delta\omega/T)\,N(\Delta\omega)$, and for scattering with a reduction in frequency (Stokes scattering) we need a factor $(\hbar\Delta\omega/T)\,[N(\Delta\omega)+1]$, where $N(\Delta\omega)$ is the Bose distribution function. If we describe Stokes scattering by using a negative $\Delta\omega$, then because of the relation $N(-\Delta\omega)+1 = -N(\Delta\omega)$ it turns out that the correcting factor for both cases is

$$\frac{\hbar\Delta\omega}{T}\,[\exp(\hbar\Delta\omega/T)-1]^{-1}. \qquad \ldots\ldots(11.5)$$

We shall thus assume $T \gg \hbar\Delta\omega$. To find the fluctuations of the distribution function we shall use a method suggested by Rytov (1953) and Landau and Lifshitz (1957 b) for the calculation of fluctuations in electrodynamics and hydrodynamics.† Using this method we shall find the fluctuation of 'the force due to chance' which we write in the kinetic equation and then, by solving this equation, we can find the fluctuations of the distribution function.

For a Fermi liquid we start from the kinetic equation, which we write in the form

$$\frac{\partial \delta n}{\partial t} + \frac{\partial \delta n}{\partial \mathbf{r}} \frac{\partial \epsilon}{\partial \mathbf{p}} - \frac{\partial n_0}{\partial \mathbf{p}} \int f(\mathbf{p},\mathbf{p}') \frac{\partial \delta n(\mathbf{p}')}{\partial \mathbf{r}}\, d\tau_{p'} = I(\delta n) + \nu(\mathbf{p},\mathbf{r},t). \qquad \ldots\ldots(11.6)$$

With the exception of the 'force due to chance' $\nu(\mathbf{p},\mathbf{r},t)$, this equation is an approximation to the kinetic equation (6.1) which is linear in δn.

† The authors thank L. P. Gor'kov, I. E. Dzialoshinskii and L. P. Pitaevskii for drawing their attention to the possibility of applying this method to the kinetic equation.

In future we shall be interested only in ranges of frequency and temperature which satisfy (11.1), i.e. for which collisions can be neglected. Thus the detailed form of the collision integral is not important, as it is only a subsidiary quantity in the calculations and will be equated to zero in the final result. Therefore we put

$$I(\delta n) = -\delta n/\tau \qquad \ldots\ldots(11.7)$$

where τ is large. Next it is necessary to find the rate of change of entropy. Noting that the number of particles and the total energy are given and using the relation

$$\delta\epsilon(\mathbf{p}) = \int f(\mathbf{p},\mathbf{p}')\,\delta n(\mathbf{p}')\,d\tau_{p'}, \qquad \ldots\ldots(11.8)$$

we find

$$\dot{S} = -\left\{\int \frac{\delta n[I(\delta n)+y]}{n_0(1-n_0)}\,d\tau_p\,dV + \frac{1}{T}\int f(\mathbf{p},\mathbf{p}')\,\delta(\mathbf{r}-\mathbf{r}')\,\delta n\,I(\delta n')\,d\tau_p\,dV\,d\tau_{p'}\,dV'\right\}. \qquad \ldots\ldots(11.9)$$

Bearing in mind that $n_0(1-n_0) \simeq T\delta(\epsilon-\mu)$, where μ is the chemical potential, it is not difficult to see that $\delta n(\mathbf{p})$ must be of the form

$$\delta n(\mathbf{p}) = \nu(\theta,\phi)\,\delta(\epsilon-\mu) \qquad \ldots\ldots(11.10)$$

where θ and ϕ are the polar angles of the vector $\mathbf{p}$. This equation expresses the fact that fluctuations of the distribution function take place only in the neighbourhood of the Fermi surface.

It is natural to assume the same form for y, i.e.

$$y(\mathbf{p}) = y^{\epsilon}(\theta,\phi)\,\delta(\epsilon-\mu). \qquad \ldots\ldots(11.11)$$

We now introduce the notation

$$F(\chi) = [f(\mathbf{p},\mathbf{p}')\,d\tau_p/d\epsilon]_{\epsilon=\epsilon'=\mu}$$

where χ is the angle between $\mathbf{p}$ and $\mathbf{p}'$ and we expand δn, y and F in spherical harmonics

$$\nu(\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} A_n^{\,m}\,\mathrm{P}_n^m(\cos\theta)\,\mathrm{e}^{im\phi},$$

$$y^{\epsilon}(\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} y_n^{\,m}\,\mathrm{P}_n^m(\cos\theta)\,\mathrm{e}^{im\phi},$$

$$F(\chi) = \sum_{n=0}^{\infty} F_n\,\mathrm{P}_n(\cos\chi). \qquad \ldots\ldots(11.12)$$

Since ν and y^{ϵ} are real

$$A_n^{\,m} = (A_n^{\,-m})^*, \quad y_n^{\,m} = (y_n^{\,-m})^*.$$

Using (11.7) for the collision integral, we get the following equation for the rate of change of entropy

$$\dot{S} = \int dV\left(\frac{d\tau_p}{d\epsilon}\right)_{\epsilon=\mu}\frac{1}{T}\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(\frac{F_n}{2n+1}+1\right)\frac{1}{2n+1}\frac{(n+|m|)!}{(n-|m|)!}\left(\frac{A_n^{\,m}}{\tau}-y_n^{\,m}\right)A_n^{\,-m}. \qquad \ldots\ldots(11.13)$$

We now introduce the notation

$$\dot{x}_n^m = -\frac{A_n^m}{\tau} + y_n^m. \qquad \ldots\ldots(11.14)$$

To put (11.13) in the form (see Landau and Lifshitz 1957 b)

$$\dot{S} = -\sum_i X_i \dot{x}_i$$

we must take as the generalized force X_i the expression

$$X_n^m = \frac{1}{T}\left(\frac{d\tau_p}{d\epsilon}\right)_{\epsilon=\mu}\left(\frac{F_n}{2n+1}+1\right)\frac{1}{2n+1}\frac{(n+|m|)!}{(n-|m|)!}A_n^{-m}. \qquad \ldots\ldots(11.15)$$

In (11.14) the quantity y_n^m plays the role of a 'force due to chance'. Writing this relation in the form

$$x_n^m = -\sum_{n'm'} \gamma_{n,n'}^{m,m'} X_{n'}^{m'} + y_n^m$$

where the coefficients γ can be found without difficulty from (11.15), we have, by the general theory of fluctuations

$$y_n^m(\mathbf{r},t)\,y_{n'}^{m'}(\mathbf{r}',t') = (\gamma_{n,n'}^{m,m'} + \gamma_{n',n}^{m',m})\,\delta(t-t')\,\delta(\mathbf{r}-\mathbf{r}')$$
$$= \frac{2}{\tau}\delta_{n,n'}\delta_{m,-m'}\,\delta(\mathbf{r}-\mathbf{r}')\,\delta(t-t')\left[T\left(\frac{d\tau_p}{d\epsilon}\right)_{\epsilon=\mu}\left(\frac{F_n}{2n+1}+1\right)\frac{1}{2n+1}\frac{(n+|m|)!}{(n-|m|)!}\right]^{-1}. \qquad \ldots\ldots(11.16)$$

Finally, using (11.12) and the relation

$$\sum_n (2n+1)\,P_n(\cos\chi) = 2\delta(\cos\chi - 1),$$

and after several transformations, we obtain the general relation

$$[y(\mathbf{p},\mathbf{r},t)\,y(\mathbf{p}',\mathbf{r}',t')]_{\mathrm{av}} = \frac{2T}{\tau}\,\delta(\mathbf{r}-\mathbf{r}')\,\delta(t-t')\left\{\frac{(2\pi\hbar)^3}{2}\,\delta(\mathbf{p}-\mathbf{p}')\,\delta(\epsilon-\mu)\right.$$
$$\left. + \delta(\epsilon-\mu)\,\delta(\epsilon'-\mu)\left(\frac{d\epsilon}{d\tau_p}\right)_{\epsilon=\mu}\sum_{n=0}^{\infty}\frac{F_n P_n(\cos\chi)}{1+F_n/(2n+1)}\right\}. \qquad \ldots\ldots(11.17)$$

From this equation and the kinetic equation (11.6) we can calculate the fluctuations in the distribution function. Since this is rather complicated in the general case for arbitrary f we shall consider f to be constant.

Noting that the fluctuations take place only at the Fermi surface and using (11.6), we express $\nu_{\mathbf{q},\Delta\omega}(\theta,\phi)$ in terms of the corresponding Fourier component of $y^{\mathbf{q}}(\theta,\phi)$ and this gives

$$\int \nu_{\mathbf{q},\Delta\omega}(\theta,\phi)\frac{d\Omega}{4\pi} = \int \frac{y_{\mathbf{q},\Delta\omega}{}^{\mathbf{q}}(\theta,\phi)\,(d\Omega/4\pi)}{-i\omega+1/\tau+i\mathbf{qv}}\Big/\left(1+F_0\int\frac{i\mathbf{qv}\,(d\Omega/4\pi)}{-i\omega+1/\tau+i\mathbf{qv}}\right) \qquad \ldots\ldots(11.18)$$

where $\mathbf{v} = (\partial\epsilon/\partial\mathbf{p})_{\epsilon=\mu}$. Taking the mean square of the modulus of this expression and using (11.17) (here it should be remembered that the Fourier component with respect to t is taken in accordance with (11.3)), we find

$$\left[\left|\int \nu_{\mathbf{q},\Delta\omega}(\theta,\phi)\frac{d\Omega}{4\pi}\right|^2\right]_{\text{av}} = 2\left(\frac{d\epsilon}{d\tau_p}\right)_{\epsilon=\mu}\frac{T}{\tau}\left[\frac{1}{2}\int_{-1}^{1}\frac{dx}{|qvx-\omega+i/\tau|^2}\right.$$
$$\left.-\frac{F_0}{1+F_0}\left|\frac{1}{2}\int_{-1}^{1}\frac{dx}{qvx-\omega+i/\tau}\right|^2\right]\Bigg/\left|1+\frac{F_0}{2}\int_{-1}^{1}\frac{qvx\,dx}{qvx-\omega+i/\tau}\right|^2. \quad \ldots\ldots(11.19)$$

We are interested in the limiting value of this expression as $\tau\to\infty$. When $qv>|\Delta\omega|$ the denominator has no poles, and the residue in the integral in the numerator becomes important. Then we find

$$\frac{1}{2\pi}[]\nu_{\mathbf{q},\Delta\omega}(\mathbf{p})\,d\tau_p|^2]_{\text{av}} = T\left(\frac{d\tau_p}{d\epsilon}\right)\frac{1}{qv}\left\{\left[1+F_0\left(1-\frac{\Delta\omega}{2qv}\ln\frac{qv+\Delta\omega}{qv-\Delta\omega}\right)\right]^2+\left(\frac{F_0\Delta\omega\pi}{2qv}\right)^2\right\}^{-1}. \quad \ldots\ldots(11.20)$$

In the other case, i.e. for $qv<\Delta\omega$, the pole in the denominator of (11.19) becomes important. It is not difficult to see that the pole arises for $F_0>0$ and equals $\Delta\omega = \pm sqv$ where s satisfies the equation

$$1+F_0\left[1-\tfrac{1}{2}s\ln\left(\frac{s+1}{s-1}\right)\right]=0. \quad \ldots\ldots(11.21)$$

This coincides with equation (9.16) which determined the velocity of zero sound. Using the relation

$$\frac{1}{\pi}\lim_{\tau\to\infty}\frac{1/\tau}{(\omega-\omega_0)^2+\tau^{-2}} = \delta(\omega-\omega_0)$$

we obtain without difficulty

$$\frac{1}{2\pi}\left[\left|\int \nu_{\mathbf{q},\Delta\omega}(\mathbf{p})\,d\tau_p\right|^2\right]_{\text{av}} = T\left(\frac{d\tau_p}{d\epsilon}\right)_{\epsilon=\mu}\frac{2(s^2-1)}{F_0(1+F_0-s^2)}[\delta(\Delta\omega-sqv)+\delta(\Delta\omega+sqv)]. \quad \ldots\ldots(11.22)$$

Thus the angular and frequency distributions of scattered light are (we have also brought in the quantum factor (11.5))

$$dh = \frac{\omega^4}{4\pi c^4}\left(\frac{\partial D}{\partial N}\right)^2\left(\frac{d\tau_p}{d\epsilon}\right)_{\epsilon=\mu}(1+\cos^2\theta)\frac{\hbar\Delta\omega}{\exp(\hbar\Delta\omega/T)-1}$$
$$\times\left[\frac{\theta(qv-|\Delta\omega|)}{2qv}\Bigg/\left\{\left[1+F_0\left(1-\frac{\Delta\omega}{2qv}\ln\frac{qv+\Delta\omega}{qv-\Delta\omega}\right)\right]^2+\left(\frac{F_0\Delta\omega\pi}{2qv}\right)^2\right\}\right.$$
$$\left.+\frac{s^2-1}{F_0(1+F_0-s^2)}\{\delta(\Delta\omega-sqv)+\delta(\Delta\omega+sqv)\}\right]\frac{d\Omega}{4\pi}\,d\Delta\omega \quad \ldots\ldots(11.23)$$

where $$\theta(y)=\begin{cases}1 & (y>0),\\ 0 & (y<0).\end{cases}$$

The result which we have obtained has a simple physical interpretation. As can easily be seen, the frequency spectrum consists of a central portion

$-qv < \Delta\omega < qv$ and two sharp lines at $\Delta\omega = \pm sqv$. The middle portion corresponds to the Doppler broadening of the initial line, and a comparison of (11.21) and (9.16) shows that the displaced lines are the satellites in Rayleigh scattering, which appear here because of the possibility of propagation of zero sound ($sv = u$). The relation between the intensities of the central line and the satellites depends in general on the angle of scattering, but in the two limiting cases of high temperatures ($T \gg \hbar\omega u/c$) and low temperatures ($T \ll \hbar\omega u/c$) this relation is independent of the angle. We can make numerical estimates for ^{3}He by using the known parameters and it turns out that at high temperatures the central portion has about 20% of the total intensity, and that the satellites have 40% each. At low temperatures the distribution is cut off for positive $\Delta\omega$ because of the quantum factor. In particular, only the Stokes satellite with $\Delta\omega = -uq$ remains of the two that were previously present, and this has 90% of the total intensity, the other 10% going to the central portion.

The total scattering intensity is obtained by integrating (11.23) over $d\Delta\omega$ and $d\Omega$. For high temperatures ($T \gg \hbar\Delta\omega \sim \hbar\omega u/c$) it equals

$$h = \frac{\omega^4 T}{6\pi c^4}\left(\frac{\partial D}{\partial N}\right)^2 \left(\frac{d\tau_p}{d\epsilon}\right)_{\epsilon=\mu} J_1 \qquad \ldots\ldots(11.24)$$

where J_1 is a numerical integral which for ^{3}He is approximately equal to 0·5. In the limit of the quantum case ($T \ll \hbar\omega u/c$) we have

$$h = \frac{\hbar\omega^5 v}{6\pi c^4}\left(\frac{\partial D}{\partial N}\right)\left(\frac{d\tau_p}{d\epsilon}\right)_{\epsilon=\mu} J_2 \qquad \ldots\ldots(11.25)$$

where J_2 is a numerical integral which is approximately equal to 0·2 for ^{3}He. To make a quantitative estimate of the results obtained for ^{3}He we need to know the value of $\partial D/\partial N$. There are no measurements of this quantity, so we assume $D-1$ is proportional to N and take the coefficient of proportionality from the data on liquid ^{4}He. For a wavelength $\lambda = 5461$ Å the refractive index of liquid ^{4}He is 1·027 and this gives $\partial D/\partial N = 2{\cdot}5 \times 10^{-24}$. Substitution in (11.24) and (11.25) gives (T in °K)

$$h(^3\text{He}) \sim 10^{-69}\,\omega^4\, T\,\text{cm}^{-1} \quad \text{for} \quad \omega \ll 2 \times 10^{17}\, T\,\text{sec}^{-1},$$

$$h(^3\text{He}) \sim 10^{-87}\,\omega^5\,\text{cm}^{-1} \quad \text{for} \quad \omega \gg 2 \times 10^{17}\, T\,\text{sec}^{-1}. \qquad \ldots\ldots(11.26)$$

We must bear in mind that the frequency must satisfy the relation (11.1), i.e. $\Delta\omega \gg 1/\tau$ or $\omega \gg 10^{18}\, T^2\,\text{sec}^{-1}$.

If this condition is not satisfied the width of the lines will be too great. Thus in the visible frequency range we need a temperature of less than 0·05°K. It is not difficult to see that at temperatures of the order of 0·01°K and frequencies in the visible range we obtain h (^{3}He) $\sim 10^{-9}\,\text{cm}^{-1}$, which is, of course, too small for the effect to be measured.† However, since ω is raised to a high power in the expression for h, it is possible that scattering will be measured successfully in the ultraviolet region.

† For liquid ^{4}He in the visible region $h \sim 10^{-8}\,\text{cm}^{-1}$ (Ginsburg 1943) (^{3}He must give approximately the same value for $\Delta\omega\tau \ll 1$). Although measurements of h have been made for ^{4}He, even they are at the limit of experimental possibilities.

APPENDIX

THE MICROSCOPIC THEORY OF A FERMI LIQUID

§ A1. A Rarefied Fermi Gas

In this section we consider the properties of a non-ideal Fermi gas, the particles of which have dimensions which are small compared with the average wavelength (Abrikosov and Khalatnikov 1957 d). This model, a series of whose properties was considered recently by Huang and Yang (1957), and also by Lee and Yang (1957), was called by them the hard sphere model. Actually it is of a more general nature and corresponds to a system of Fermi particles with arbitrary short-range forces whose range of action is small compared with the average wavelength.

Such a model enables us to obtain expressions for the various quantities which describe a Fermi liquid, for example the energy, the effective mass and the function f as a power series in $a/\lambda\!\!\!^{-}$ (a is the radius and $\lambda\!\!\!^{-}$ the wavelength). Our calculation neglects terms of higher order than $(a/\lambda\!\!\!^{-})^2$. In principle it would be possible to determine the terms of several subsequent orders in an analogous way, but this is of no special interest. We begin with a calculation of the energy.

We use perturbation theory for the interaction energy of particles, which we write in the form (taking for simplicity unit volume of the gas)

$$V = 2U \sum_{\substack{n_1 n_2 m_1 m_2 \\ n_1 < n_2}} a_{m_1}^{+} a_{m_2}^{+} a_{n_2} a_{n_1} \qquad \ldots\ldots\text{(A1.1)}$$

where a_i^{+} and a_i are the particle creation and annihilation operators. The summation is made taking into account conservation of the total momentum and that the projection of the spins of the states m_1 and n_1 are equal, and similarly for m_2 and n_2. The removal of U from under the summation sign corresponds to the fact that the interaction between any of the pairs of particles is the same, and that the scattering amplitude is independent of the angle. As a first approximation U is related to this amplitude by the expression

$$U = 4\pi a\hbar^2/m \qquad \ldots\ldots\text{(A1.2)}$$

(a is the amplitude for s-scattering).

The first-order correction to the energy equals the diagonal matrix element of V

$$E_n^{(1)} = 2U \sum_{n_1 < n_2} N_{n_1} N_{n_2} Q_{n_1 n_2} \qquad \ldots\ldots\text{(A1.3)}$$

where the N_i are occupation numbers.

The factor Q_{ik} in (A1.3) takes into account the fact that Fermi particles whose scattering amplitude is independent of angle do not interact when their spins are parallel. Thus we shall assume that

$$Q_{ik} = \tfrac{1}{4} - \sigma_i \sigma_k \qquad \ldots\ldots\text{(A1.4)}$$

(σ_i is the spin operator of the ith particle).

Substituting (A1.2) and (A1.4) into (A1.3), we get

$$E^{(1)} = (2\pi a\hbar^2/m) N^2/2. \qquad \ldots\ldots\text{(A1.5)}$$

To find the second-order correction we use the perturbation theory expression

$$E_n^{(2)} = \sum_{m \neq n} |V_{nm}|^2/(E_n - E_m). \qquad \ldots\ldots(A1.6)$$

Substituting (A1.1) in this expression we get the following sum

$$4U^2 \sum_{\substack{n_1 n_2 m_1 m_2 \\ n_1 < n_2}} \frac{N_{n_1} N_{n_2}(1-N_{m_1})(1-N_{m_2}) Q_{n_1 n_2}}{E_{n_1} + E_{n_2} - E_{m_1} - E_{m_2}} \qquad \ldots\ldots(A1.7')$$

where the N_i are equilibrium occupation numbers, and the E_i are the energies of the particles.

Since our aim is to obtain an expansion for the energy in powers of a, we must bear in mind that the relation (A1.2) between U and the scattering amplitude is not exact and is accurate only to first-order terms. Taking into account the second-order terms we get the following equation instead of (A1.2)

$$2U + 4U^2 \sum_{m_1 m_2} \frac{Q_{n_1 n_2}}{E_{n_1} + E_{n_2} - E_{m_1} - E_{m_2}} = \frac{8\pi a \hbar^2}{m}. \qquad \ldots\ldots(A1.2')$$

If we now express U in terms of a and substitute the result in (A1.3), the expression for $E^{(1)}$ will contain terms in a^2 which should naturally be added to the second-order correction. Taking this into account we obtain the following value of the second approximation to the energy

$$E^{(2)} = 2U^2 \sum_{n_1 n_2 m_1 m_2} \left\{ \frac{N_{n_1} N_{n_2}(1-N_{m_1})(1-N_{m_2}) Q_{n_1 n_2}}{E_{n_1} + E_{n_2} - E_{m_1} - E_{m_2}} - \frac{N_{n_1} N_{n_2} Q_{n_1 n_2}}{E_{n_1} + E_{n_2} - E_{m_1} - E_{m_2}} \right\}. \qquad \ldots\ldots(A1.7)$$

Since the expression in brackets is symmetrical in n_1 and n_2 we replace the condition $n_1 > n_2$ by the factor $\frac{1}{2}$.

The point of this method is that an expansion in powers of U is not valid here. The presence of a constant U would simply lead to an infinity in the energy, as can be seen directly from (A1.7'). In this case it is important that the scattering amplitude a is finite and, moreover, small, which enables us to make a power series expansion in terms of a.

In the first part of the expression (A1.7) the term containing the product of the four N_i is zero because the denominator is antisymmetric with respect to the interchange $n_1 n_2 \to m_1 m_2$, while the numerator is symmetric, and all the regions of the summation are identical. The remaining two terms, containing the product of three N_i are equal. Thus we finally obtain

$$E^{(2)} = -4U^2 \sum_{n_1 n_2 m_1 m_2} \frac{N_{n_1} N_{n_2} N_{m_2} Q_{n_1 n_2}}{E_{n_1} + E_{n_2} - E_{m_1} - E_{m_2}}. \qquad \ldots\ldots(A1.8)$$

This is the basic expression for the energy and is valid provided the initial condition $(a/\lambda \ll 1)$ is fulfilled.

Our aim is to obtain the properties of a degenerate Fermi gas. From (A1.8) we find the second order correction to the energy of the ground state

$$E^{(2)} = -\frac{4U^2}{(2\pi\hbar)^6}\int d\mathbf{p}_1\int d\mathbf{p}_2\int d\mathbf{p}_3\int d\mathbf{p}_4\,\frac{\delta(\mathbf{p}_1+\mathbf{p}_2-\mathbf{p}_3-\mathbf{p}_4)}{(\mathbf{p}_1^2+\mathbf{p}_2^2-\mathbf{p}_3^2-\mathbf{p}_4^2)/2m}, \qquad \text{.....(A1.9)}$$

$$|\mathbf{p}_1|<p_0,\ |\mathbf{p}_2|<p_0,\ |\mathbf{p}_3|<p_0,$$

where p_0 is the limiting momentum.

By § 2, the energy of excitations is determined by the relation

$$\epsilon_i = \delta E/\delta N_i. \qquad \text{.....(A1.10)}$$

Variation of equations (A1.3) and (A1.8) with respect to N_i gives

$$\epsilon(\mathbf{p}) = \frac{\mathbf{p}^2}{2m}+\frac{UN}{2}+\frac{2U^2}{(2\pi\hbar)^6}\int d\mathbf{p}_1\int d\mathbf{p}_2\int d\mathbf{p}_3\Big[\frac{\delta(\mathbf{p}_1+\mathbf{p}_2-\mathbf{p}-\mathbf{p}_3)}{(\mathbf{p}^2+\mathbf{p}_3^2-\mathbf{p}_1^2-\mathbf{p}_2^2)/2m}$$

$$-2\,\frac{\delta(\mathbf{p}_1+\mathbf{p}-\mathbf{p}_2-\mathbf{p}_3)}{(\mathbf{p}^2+\mathbf{p}_1^2-\mathbf{p}_2^2-\mathbf{p}_3^2)/2m}\Big]. \qquad \text{.....(A1.11)}$$

$$|\mathbf{p}_1|<p_0,\ |\mathbf{p}_2|<p_0,$$

Thus the problem of the calculation of the energy of the ground state and of the effective mass of the excitations reduces to the evaluation of the integrals (A1.9) and (A1.11). The integration is rather unwieldy in view of the multiple integral and the awkward region of integration.

Instead of this we can apply a simpler method, based on the introduction of the function f. If we introduce this function

$$f_{ik} = \delta^2 E/\delta N_i\,\delta N_k \qquad \text{.....(A1.12)}$$

then, by (3.5) and (9.7), we can determine the effective mass and the velocity of low-frequency sound. The energy of the ground state can be found from the velocity of sound by integration.

Thus the problem reduces to the determination of the quantity f. By varying (A1.3) and (A1.8) with respect to N_i and then N_k, we find the following expression :

$$f = 2UQ_{\sigma,\sigma'}-\frac{8U^2}{(2\pi\hbar)^3}\int d\mathbf{p}_1\int d\mathbf{p}_2\Big[Q_{\sigma,\sigma'}\,\frac{\delta(\mathbf{p}+\mathbf{p}'-\mathbf{p}_1-\mathbf{p}_2)}{(\mathbf{p}^2+\mathbf{p}'^2-\mathbf{p}_1^2-\mathbf{p}_2^2)/2m}$$

$$+\frac{1}{4}\,\frac{\delta(\mathbf{p}+\mathbf{p}_1-\mathbf{p}'-\mathbf{p}_2)}{(\mathbf{p}^2+\mathbf{p}_1^2-\mathbf{p}'^2-\mathbf{p}_2^2)/2m}+\frac{1}{4}\,\frac{\delta(\mathbf{p}'+\mathbf{p}_1-\mathbf{p}-\mathbf{p}_2)}{(\mathbf{p}'^2+\mathbf{p}_1^2-\mathbf{p}^2-\mathbf{p}_2^2)/2m}\Big]. \qquad \text{.....(A1.13)}$$

$$|\mathbf{p}_1|<p_0,$$

In the calculation we shall immediately put $|\mathbf{p}|=|\mathbf{p}'|=p_0$. The integration of the second term in f is considerably simpler than the integrals (A1.9) and (A1.11). As a result we find

$$f(\chi) = \frac{2\pi a\hbar^2}{m}\Big[1+2\Big(\frac{3}{\pi}\Big)^{1/3}aN^{1/3}\Big(2+\frac{\cos\chi}{2\sin\frac{1}{2}\chi}\ln\frac{1+\sin\frac{1}{2}\chi}{1-\sin\frac{1}{2}\chi}\Big)\Big]$$

$$-\frac{8\pi a\hbar^2}{m}(\boldsymbol{\sigma}_1\boldsymbol{\sigma}_2)\Big[1+2\Big(\frac{3}{\pi}\Big)^{1/3}aN^{1/3}\Big(1-\frac{\sin\frac{1}{2}\chi}{2}\ln\frac{1+\sin\frac{1}{2}\chi}{1-\sin\frac{1}{2}\chi}\Big)\Big]. \qquad \text{.....(A1.14)}$$

This expression has a feature which is worthy of note: that for angles χ near to π the function f has a logarithmic singularity

$$f(\chi) \sim (\tfrac{1}{4} - \boldsymbol{\sigma}_1 \boldsymbol{\sigma}_2) \ln \frac{1}{\pi - \chi}. \qquad \ldots\ldots\text{(A1.15)}$$

It is clear that in this case the approximation we have used is not strictly applicable. The singularity of the function f at $\chi = \pi$ is a reflection of the singularity in the scattering amplitude for excitations which collide at an angle π. The corresponding expression obtained by summing the main terms given by perturbation theory is proportional to

$$\left[1 + 2\left(\frac{3}{\pi}\right)^{1/3} aN^{1/3}\left(\ln \frac{p_0^2}{\epsilon} + i\frac{\pi}{2}\right)\right]^{-1} \qquad \ldots\ldots\text{(A1.16)}$$

where $\epsilon = \mathbf{p}^2 + \mathbf{p}'^2 - 2p_0^2$.

For a positive, this expression reduces to zero for $p^2 = p'^2 = p_0^2$.

If $a < 0$ (this is permissible for a Fermi system), the scattering amplitude has a pole at a small imaginary value of ϵ. This corresponds to the instability of the ground state with respect to the formation of bound pairs from excitations of opposite momenta, as was noted by Cooper (1956), and is the cause of the phenomenon of superconductivity in metals (Bardeen, Cooper and Schrieffer 1957).

Thus the expression obtained for f is invalid for angles near to π. However, since the singularity is logarithmic, it affects only the immediate neighbourhood of the singular point. Because the expressions which interest us contain only integrals of f with regular functions, the logarithmic singularity of f is not important.

Substituting (A1.4) into (3.5) we find the value of the effective mass

$$\frac{m}{m^*} = 1 - \frac{8}{15}\left(\frac{3}{\pi}\right)^{2/3}(7\ln 2 - 1). \qquad \ldots\ldots\text{(A1.17)}$$

We note that if we substitute in here the value $m^* = 1{\cdot}43m$, which corresponds to liquid helium 3, then we get a value of a of $1{\cdot}6 \times 10^{-8}$ cm, i.e. of the order of the diameter of the helium atom in gas kinetics. This comparison has, of course, no real meaning and the model considered cannot describe liquid ^{3}He. This is evident from the fact that the quantity $(m^* - m)/m^*$, which should, according to the theory, be second order in a, is in fact equal to 1/3 for ^{3}He.

Substituting the formula for f in the expression for the velocity of sound, we get

$$c^2 = \frac{\pi^{4/3}}{3^{1/3}} N^{2/3} \frac{\hbar^2}{m^2} + 2\frac{\pi a \hbar^2}{m^2} N\left[1 + \frac{4}{15}\left(\frac{3}{\pi}\right)^{1/3} aN^{1/3}(11 - 2\ln 2)\right]. \qquad \ldots\ldots\text{(A1.18)}$$

From the value obtained for c^2 it is not difficult to obtain the energy of the ground state of a Fermi liquid. We use the relation (9.2), $c^2 = (N/m)\partial\mu/\partial N$, and this gives

$$E = \int \mu dN = E^{(0)} + \frac{\pi a \hbar^2}{m} N^2\left[1 + \frac{6}{35}\left(\frac{3}{\pi}\right)^{1/3} aN^{1/3}(11 - 2\ln 2)\right]. \qquad \ldots\ldots\text{(A1.19)}$$

Formula (A1.19) coincides with the result of Lee and Yang (1957).

The same result may be obtained by a direct integration of (A1.9). In the evaluation of this integral it is convenient to introduce the new variables

$$\mathbf{p} = \mathbf{p}_1 - \mathbf{p}_2, \quad \mathbf{q} = \mathbf{p}_3 - \mathbf{p}_4, \quad \mathbf{s} = \mathbf{p}_1 + \mathbf{p}_2 = \mathbf{p}_3 + \mathbf{p}_4.$$

In these variables $E^{(2)}$ becomes

$$E^{(2)} = -\frac{mU^2}{4(2\pi\hbar)^2}\int d\mathbf{s}\int d\mathbf{p}\int d\mathbf{q}\,\frac{1}{p^2-q^2}$$

where the region of integration for the vector $\mathbf{s}$ is $0<|\mathbf{s}|<2p_0$, and the range of values of $\mathbf{p}$ and $\mathbf{q}$ is shown in figure 2. If we introduce the variables $x_q = \cos(\mathbf{q},\mathbf{s})$

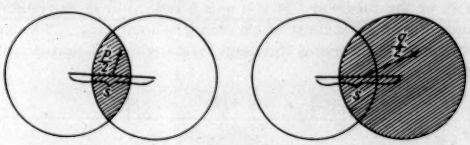


Figure 2.

and $x_p = \cos(\mathbf{p},\mathbf{s})$, then $E^{(2)}$ becomes

$$E^{(2)} = -\frac{8mU^2\pi^3}{(2\pi\hbar)^9}\int_0^{2p_0} s^2\,ds\int_0^1 dx_p\int_0^{z(x_p)} p^2\,dp\int_{-1}^1 dx_q\int_0^{z(x_q)} q^2\,dq\,\frac{1}{p^2-q^2}$$

where $z(x)$ satisfies the equation

$$z^2+2zsx+s^2 = 4p_0^2.$$

By a series of transformations and integrations by parts with respect to dx_p and dx_q we find

$$E^{(2)} = \frac{2mU^2p_0^7}{\pi^6\hbar^9}\int_0^1 s^2\,ds\left[\int_0^{1+s} p^2\,dp\int_0^{1-s} q^2\,dq\right.$$
$$\left.+\frac{1}{4s^2}\int_{\sqrt{(1-s^2)}}^{1+s} p\,dp(1-p^2-s^2)\int_{\sqrt{(1-s^2)}}^{1-s} q\,dq(1-q^2-s^2)\right]\frac{1}{p^2-q^2}.$$

A further integration by parts with respect to s and evaluation of the remaining integrals gives

$$E^{(2)} = \frac{6}{35}\left(\frac{3}{\pi}\right)^{1/3}(11-2\ln 2)\,aN^{1/3}E^{(1)}.$$

Here we have expressed U by formula (A1.2) and substituted $p_0 = \hbar(3\pi^2N)^{1/3}$. This result coincides with the second-order term in (A1.19).

§ A2. The Microscopic Theory of a Fermi Liquid at $T = 0$

Here we give an account of the general microscopic theory of a Fermi liquid at $T = 0$ which was recently developed by Landau (1958), and which is valid for an arbitrary interaction. As we shall see, the result of this treatment will provide a basis for a phenomenological theory and enable us to understand the physical meaning of the function f which appears in it.

The microscopic treatment is based on methods taken from quantum field theory, which, as is known, operates with products of field operators which are taken in chronological order and averaged over the vacuum. These are the so-called Green's functions. In our case we shall consider averages over the ground state of

the system instead of over a vacuum and it is not difficult to see that these averages have the same properties as the Green's functions of field theory and, in particular, may in principle be calculated using the technique of the Feynman graphs.

To be explicit, we shall call a Green's function the quantity

$$G_{1,2} = -i\langle T(\psi_1 \psi_2^+)\rangle \qquad \ldots\ldots(A2.1)$$

where the brackets $\langle\,\rangle$ denote an average over the ground state, ψ and ψ^+ are operators of the Fermi particle field, the indices 1, 2 denote the totality of space–time and spin coordinates and, as usual, T denotes a chronological product (with a change of sign when ψ_1 and ψ_2^+ are interchanged). In the absence of an external field the Green's function depends only on the difference in coordinates and times. As far as the spin indices are concerned, if the particle interaction is spin-independent, then $G_{\alpha\beta} \sim \delta_{\alpha\beta}$.

The Fourier representation of the Green's function of the coordinate and time difference will be denoted by $G(P)\,\delta_{\alpha\beta}$ (here P is the four-dimensional vector $(\mathbf{p}, \epsilon)$). The pole of $G(P)$ determines the energy of a quasi-particle. If the modulus of $\mathbf{p}$ is near to p_0 and ϵ is near to the boundary energy μ, then $G(P)$ becomes

$$G(P) = \frac{a}{\epsilon - \mu - v(p - p_0) + i\delta}. \qquad \ldots\ldots(A2.2)$$

Here a is a positive constant, δ is a small quantity whose sign is the same as that of $\epsilon - \mu$ or, near the pole, as that of $p - p_0$ (it is introduced to define the rule for going round poles, which corresponds to taking the T-product in the $(\mathbf{x}, t)$ representation).

We now introduce the average of the chronological product of the four operators

$$\Phi_{12,34} = \langle T(\psi_1 \psi_2 \psi_3^+ \psi_4^+)\rangle. \qquad \ldots\ldots(A2.3)$$

We call the quantity $\Gamma_{12,34}$ the vertex part, and it is defined by the following relation

$$\Phi_{12,34} = G_{1,3} G_{2,4} - G_{1,4} G_{2,3} + i \sum_{1',2',3',4'} G_{1,1'} G_{2,2'} G_{3,3'} G_{4,4'} \Gamma_{1'2',3'4'}. \qquad \ldots\ldots(A2.4)$$

The Fourier representation of the vertex part is

$$(2\pi)^4 \Gamma_{\alpha\beta,\gamma\delta}(P_1 P_2, P_3 P_4)\,\delta(P_1 + P_2 - P_3 - P_4).$$

The function $\Gamma_{\alpha\beta,\gamma\delta}(P_1 P_2, P_3 P_4)$ evidently changes sign for the interchange $\alpha P_1 \rightleftharpoons \beta P_2$. Analogously to quantum field theory, when there is a definite relation between the components of the vector P (for example, near the Fermi surface $\epsilon - \mu = v(p - p_0)$) the function Γ multiplied by a^2 plays the role of the scattering amplitude of quasi-particles.

We shall consider Γ in the case when P_1 is near to P_3 and therefore P_2 is near to P_4, and we introduce the definition

$$\Gamma(P_1 P_2, P_1 + K, P_2 - K) \equiv \Gamma(P_1 P_2;\, K)$$

where $K = (\mathbf{K}, \omega)$ is a small four-vector.

If such a vertex part could be calculated by perturbation theory, then for the first order it would be necessary to take the graphs given in figure 3 (I). The

corresponding expressions contain integrals of the product of two Green's functions. While the two graphs (*a*) and (*b*) are regular at $K = 0$, for (*c*) the poles of the two Green's functions come together when $K = 0$, which, as will be seen below, leads to a singularity in Γ.

If we denote by $\Gamma^{(1)}$ the totality of all possible graphs for Γ which have no 'peculiar elements' (i.e. integrals of $G(P)G(P+K)$) it is not difficult to see that the total Γ is obtained by the summation of the 'ladder', given in figure 3 (II).

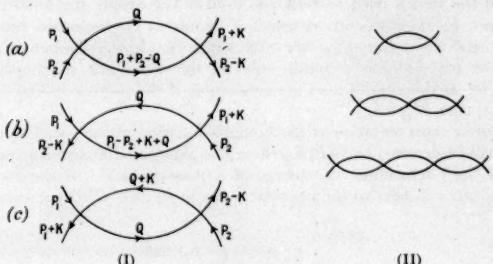


Figure 3.

This can be carried out by the solution of the integral equation

$$\Gamma_{\alpha\beta,\gamma\delta}(P_1 P_2; K) = \Gamma_{\alpha\beta,\gamma\delta}{}^{(1)}(P_1 P_2) - \frac{i}{(2\pi\hbar)^4}\int \Gamma_{\alpha\epsilon,\gamma\zeta}{}^{(1)}(P_1 Q)\, G(Q)\, G(Q+K) \times \Gamma_{\zeta\beta,\epsilon\delta}(Q P_2; K)\, d^4 Q. \qquad \ldots\ldots (A2.5)$$

Here, since $\Gamma^{(1)}$ has no singularities for $K = 0$, we simply put $K = 0$ everywhere in $\Gamma^{(1)}$.

We now consider the product $G(Q)\,G(Q+K)$ in the integral. The integral with respect to Q will consist of a main value and a term which comes from the poles. Since the arguments of both G are approximately the same, we can assume that all the other terms in the integrals at these poles have the same value. Thus there is a contribution from the poles only when they are on different sides of the real axis. This implies $q < p_0$, $|\mathbf{q}+\mathbf{k}| > p_0$, or vice versa. Taking into account the smallness of $\mathbf{k}$ and ω it is not difficult to see that

$$q \simeq p_0 \quad \text{and} \quad \epsilon_q \simeq \mu.$$

Thus in the part of the integral with respect to Q which is connected with the poles, the product $G(Q)\,G(Q+K)$ can be replaced by

$$A\delta(\epsilon-\mu)\,\delta(q-p_0).$$

The coefficient A can be determined by the integration of $G(Q)\,G(Q+K)$ and is $2\pi i a^2 k\cos\theta/(\omega - vk\cos\theta)$ where θ is the angle between $\mathbf{q}$ and $\mathbf{k}$. Thus the product $G(Q)\,G(Q+K)$ can be written as

$$G(Q)\,G(Q+K) = 2\pi i a^2 \frac{k\cos\theta}{\omega - vk\cos\theta}\,\delta(\epsilon-\mu)\,\delta(q-p_0) + \phi(Q) \qquad \ldots\ldots (A2.6)$$

where ϕ corresponds to the main value of the integral and has no singularities (therefore it can be assumed that $K = 0$).

If we consider the expression (A2.6) as K tends to 0, it is very important how this limit is taken, i.e. what relation we take between ω and k. The same is true for Γ in the limit $\omega \to 0$, $\mathbf{k} \to 0$.

First we shall consider Γ in the limit $\mathbf{k} \to 0$, $k/\omega \to 0$. For this limit, which we write as Γ^{ω}, we get from (A2.5) and (A2.6):

$$\Gamma_{\alpha\beta,\gamma\delta}{}^{\omega}(P_1 P_2) = \Gamma_{\alpha\beta,\gamma\delta}{}^{(1)}(P_1 P_2) - \frac{i}{(2\pi\hbar)^4}\int \Gamma_{\alpha\xi,\gamma\zeta}{}^{(1)}(P_1 Q)\,\phi(Q)\,\Gamma_{\zeta\beta,\epsilon\delta}{}^{\omega}(QP_2)\,d^4Q. \quad \ldots\ldots(A2.7)$$

It is not difficult to eliminate $\Gamma^{(1)}$ from the two equations (A2.5) and (A2.7). Using (A2.6) we find on simplification

$$\Gamma_{\alpha\beta,\gamma\delta}(P_1 P_2; K) = \Gamma_{\alpha\beta,\gamma\delta}{}^{\omega}(P_1 P_2) + \frac{a^2 p_0{}^2}{(2\pi\hbar)^3}\int \Gamma_{\alpha\epsilon,\gamma\zeta}{}^{\omega}(P_1 Q) \times \Gamma_{\zeta\beta,\epsilon\delta}(QP_2, K)\,\frac{k\cos\theta}{\omega - vk\cos\theta}\,d\Omega. \quad \ldots\ldots(A2.8)$$

We now take the other limit, namely $\mathbf{k} \to 0, \omega/k \to 0$. This limiting value multiplied by a^2 corresponds to the forward physical scattering amplitude for a quasi-particle with an energy at the Fermi boundary. We write this quantity as $\Gamma^{\mathbf{K}}$. From (A2.8) we find a relation between $a^2\,\Gamma^{\mathbf{K}}$ and $a^2\,\Gamma^{\omega}$

$$a^2\,\Gamma_{\alpha\beta,\gamma\delta}{}^{\mathbf{K}}(P_1 P_2) = a^2\,\Gamma_{\alpha\beta,\gamma\delta}{}^{\omega}(P_1 P_2) - \frac{p_0{}^2}{v(2\pi\hbar)^3}\int a^2\,\Gamma_{\alpha\epsilon,\gamma\zeta}{}^{\omega}(P_1 Q)\,a^2\Gamma_{\zeta\beta,\epsilon\delta}{}^{\mathbf{K}}(QP_2)\,d\Omega. \quad \ldots\ldots(A2.9)$$

We now investigate the poles of the function $\Gamma(P_1 P_2; K)$. Similarly to the way in which the poles of the function $G(P)$ give the relation between theory and the momentum of Fermi excitations, the poles of the function Γ characterize the 'two-particle' or Bose excitations. Since $\Gamma(P_1 P_2; K) \gg \Gamma^{\omega}(P_1, P_2)$ in the neighbourhood of a pole, we can neglect Γ^{ω} on the right-hand side. It may be noted that the variable P_2 and the indices β and δ are parameters in the equations. The function Γ can be written as the product $\chi_{\alpha\gamma}(P_1; K)\,\psi_{\beta\delta}(P_2)$ and $\psi(P_2)$ cancels out on both sides of the equation. If we write

$$\nu_{\alpha\gamma}(\mathbf{n}) = \frac{\mathbf{nk}}{\omega - v\mathbf{nk}}\,\chi_{\alpha\gamma}(P_1; K)$$

where $\mathbf{n}$ is a unit vector in the direction of $\mathbf{p}_1$, then for $\nu_{\alpha\gamma}$ we get the equation

$$(\omega - v\mathbf{nk})\,\nu_{\alpha\gamma}(\mathbf{n}) = \mathbf{nk}\,\frac{p_0{}^3 a^2}{(2\pi\hbar)^3}\int \Gamma_{\alpha\epsilon,\gamma\zeta}{}^{\omega}(\mathbf{n}, \mathbf{l})\,\nu_{\zeta\epsilon}(\mathbf{l})\,d\Omega. \quad \ldots\ldots(A2.10)$$

Comparing this equation with (9.14) and (11.4) it is not difficult to see that it coincides with the equation for zero sound and spin waves, and that the quantity

$a^2\Gamma^{\omega}$ plays the role of the function f. This confirms the equation for zero sound which was obtained from the phenomenological theory and also the assumption made in the latter of the functional dependence of the energy of excitations on the distribution function. On the other hand, equation (A2.9) establishes a relation between the function f and the scattering amplitude of a quasi-particle at an angle of 0°.

We shall now consider this relation in more detail. We write the amplitude, multiplied by $(d\tau/d\epsilon)_{\epsilon=\mu}$ as $A(\mathbf{n}_1\sigma_1, \mathbf{n}_2\sigma_2)$ and we get

$$A(\mathbf{n}_1\sigma_1, \mathbf{n}_2\sigma_2) = \Phi(\mathbf{n}_1\sigma_1, \mathbf{n}_2\sigma_2) - \tfrac{1}{2}\mathrm{Sp}_{\sigma'}\int \Phi(\mathbf{n}_1\sigma_1, \mathbf{n}'\sigma')\, A(\mathbf{n}'\sigma', \mathbf{n}_2\sigma_2)\frac{d\Omega}{4\pi}. \quad \ldots\ldots(A2.11)$$

As regards spin dependence, both Φ and A contain two terms: one spin-independent and the other proportional to $\boldsymbol{\sigma}_1\boldsymbol{\sigma}_2$. If we write

$$A(\mathbf{n}_1\sigma_1, \mathbf{n}_2\sigma_2) = B(\mathbf{n}_1, \mathbf{n}_2) + \boldsymbol{\sigma}_1\boldsymbol{\sigma}_2\, C(\mathbf{n}_1, \mathbf{n}_2), \quad \ldots\ldots(A2.12)$$

then B and C satisfy the equations

$$B(\mathbf{n}_1, \mathbf{n}_2) = F(\mathbf{n}_1, \mathbf{n}_2) - \int F(\mathbf{n}_1, \mathbf{n}')\, B(\mathbf{n}', \mathbf{n}_2)\frac{d\Omega}{4\pi},$$

$$C(\mathbf{n}_1, \mathbf{n}_2) = Z(\mathbf{n}_1, \mathbf{n}_2) - \tfrac{1}{4}\int Z(\mathbf{n}_1, \mathbf{n}')\, C(\mathbf{n}', \mathbf{n}_2)\frac{d\Omega}{4\pi}. \quad \ldots\ldots(A2.13)$$

Being scalars, the quantities B and C, like F and Z, can depend only on $\cos\chi$. If we expand in spherical harmonics, $B(\chi) = B_l P_l(\cos\chi)$, etc., then we can easily find the relations between the coefficients B_l and F_l, C_l and Z_l

$$B_l = \frac{F_l}{1 + F_l/(2l+1)}, \qquad C_l = \frac{Z_l}{1 + Z_l/(2l+1)}. \quad \ldots\ldots(A2.14)$$

As an example we shall find the scattering amplitude of excitations for a rarefied Fermi gas. From (A1.16) and (A2.13) we find

$$A(\chi, \sigma_1\sigma_2) = 8\left(\frac{3}{\pi}\right)^{1/3} aN^{1/3}(\tfrac{1}{4} - \boldsymbol{\sigma}_1\boldsymbol{\sigma}_2)$$
$$+ 16\left[\left(\frac{3}{\pi}\right)^{1/3} aN^{1/3}\right]^2\left[\tfrac{1}{4} - \boldsymbol{\sigma}_1\boldsymbol{\sigma}_2 + \frac{\cos\chi}{8\sin\frac{1}{2}\chi}\ln\frac{1+\sin\frac{1}{2}\chi}{1-\sin\frac{1}{2}\chi}\right.$$
$$\left. - \boldsymbol{\sigma}_1\boldsymbol{\sigma}_2\left(1 - \frac{\sin\frac{1}{2}\chi}{2}\ln\frac{1+\sin\frac{1}{2}\chi}{1-\sin\frac{1}{2}\chi}\right)\right]. \quad \ldots\ldots(A2.15)$$

We note that for parallel spins this reduces to zero for $\chi = 0$ as it must to conform with the Pauli principle (this does not hold for f).

References

Abraham, B., Osborne, D., and Weinstock, B., 1955, *Phys. Rev.*, **98**, 551.

Abrikosov, A. A., and Khalatnikov, I. M., 1957 a, *J. Exp. Theor. Phys.*, **32**, 915; 1957 b, *Ibid.*, **32**, 1084; 1957 c, *Ibid.*, **33**, 110; 1957 d, *Ibid.*, **33**, 1154; 1958, *Ibid.*, **34**, 198.

Bardeen, J., Cooper, L., and Schrieffer, J., 1957, *Phys. Rev.*, **106**, 162.

Cooper, L., 1956, *Phys. Rev.*, **104**, 1189.

Fairbank, W., Ard, W., and Walters, G., 1954, *Phys. Rev.*, **95**, 566.

Ginsburg, V. L., 1943, *J. Exp. Theor. Phys.*, **13**, 243.

Huang, K., and Yang, C. N., 1957, *Phys. Rev.*, **105**, 767.

Lagner, H. L., Sydoriak, S. G., and Roberts, T. R., 1957, *Program of the Fifth Int. Conf. on Low Temp. Phys. and Chem.*, Univ. Wisconsin, 41.

Landau, L. D., 1956, *J. Exp. Theor. Phys.*, **30**, 1058; 1957, *Ibid.*, **32**, 59; 1958, *Ibid.*, **34**, 262.

Landau, L. D., and Lifshitz, E. M., 1957 a, *The Electrodynamics of Continuous Media.* (In Russian.) Moscow GTTI; 1957 b, *J. Exp. Theor. Phys.*, **32**, 618.

Lee, T. D., and Yang, C. N., 1957, *Phys. Rev.*, **105**, 1119.

Pomeranchuk, I. Ia., 1950, *J. Exp. Theor. Phys.*, **20**, 919.

Rytov, S. M., 1953, *The Theory of Electrical Fluctuations and Thermal Radiation.* (In Russian.) Izd. Akad. Nauk S.S.S.R.

Zinov'eva, K. N., 1958, *J. Exp. Theor. Phys.*, **34**, 609.

INVARIANCE IN ELEMENTARY PARTICLE PHYSICS

By N. KEMMER, J. C. POLKINGHORNE† and D. L. PURSEY
Tait Institute of Mathematical Physics, University of Edinburgh

CONTENTS

	PAGE
§ 1. Introduction	369
§ 2. Invariance principles and conservation laws	371
§ 3. Rotations and Lorentz transformations	374
3.1. Rotations in three dimensions	374
3.2. Rotations in four dimensions	376
3.3. Lorentz transformations	378
§ 4. Gauge transformations	380
4.1. Gauge transformations in electrodynamics	380
4.2. Conservation of baryons and leptons	384
§ 5. Space inversion, charge conjugation and time reversal	385
5.1. Space inversion	385
5.2. Charge conjugation	390
5.3. Time reversal	395
5.4. The *CPT* theorem	400
5.5. Tests of invariance, and determination of intrinsic parities	402
§ 6. Isobaric spin (nucleons and pions)	407
6.1. Nucleons	407
6.2. Pions and nucleons	410
6.3. Invariance of interaction	412
6.4. Application to scattering	414
§ 7. Isobaric spin (strange particles)	416
7.1. Associated production and the Gell-Mann–Nishijima scheme	416
7.2. Mathematical formulations of the Gell-Mann–Nishijima scheme	421
7.3. Global symmetry	423
7.4. Electric charge	426
§ 8. Conclusion	427
Appendix : Form invariance	428

Abstract. An account, as non-technical as possible, is given of the invariance properties that have important applications in elementary particle physics. These can be classified in several ways. Firstly some are related to continuous groups (rotations), others to discrete groups (inversions). Secondly some are related to groups of well-known geometrical or kinematical significance (ordinary space time), others are not and are peculiar to elementary particle physics (isobaric space). Finally some are, it appears, rigorously valid while others are of approximate nature. The discussion aims to cover all cases and their interpretation. In as far as possible explicit use of quantum field theory techniques is avoided.

---

† Now at Trinity College, Cambridge.

§ 1. Introduction

The concept of invariance is closely related to that of symmetry. It is a familiar fact, even at a very elementary level of physical theory, that the exploitation of symmetry properties is a very powerful device in solving physical problems.

The symmetries of a problem may be accidental and peculiar to it—e.g. the gravitational field of a massive cube—but of more importance are those symmetries which have their root in the inherent nature of the physical laws involved. The existence of such fundamental symmetries is closely related to the validity of general physical conservation laws and it is this fact which makes the fundamental symmetry principles so important in modern theoretical physics.

In § 2 of this article the general relation of symmetry principles and conservation laws is discussed.

The most obvious and at the same time one of the most important fundamental symmetries is that which we employ when we describe physical processes as taking place ' in three-dimensional Euclidean space '. This implies, among other things, that in the face of much contrary evidence from elementary statics and dynamics we assert that the ' up–down ' direction is equivalent to the ' backwards–forwards ' and ' right–left ' directions or, in fact, to any other direction we may specify. How the symmetries of Euclidean space lead one to a formulation of physical laws in the language of vector or tensor calculus is too well known to require elaboration, but it would be wrong to assume that with the establishment of such a calculus the exploitation of the symmetries of space is completed. With the development of Quantum Theory a deeper aspect of the self-same symmetries, namely the mathematical theory of the rotation group in three dimensions, has become an increasingly powerful tool of the theoretical physicist and although today this theory also is very generally known and used in physics, a brief sketch of it is the natural beginning of a survey of invariance in physical theory. Such a sketch will be given below in § 3.1.

The Special Theory of Relativity has taught us that the symmetries which are a result of the properties of three-dimensional space form part of a wider system of symmetries, the existence of which is expressed by the statement that physical events are to be described as occurring in the ' four-dimensional space–time continuum '. The notion of a (three-dimensional) rotation then generalizes to that of a (four-dimensional) ' Lorentz transformation '. §§ 3.2 and 3.3 deal with the theory of the Lorentz group on which this generalization depends. These sections are, of necessity, also a sketch, in this case not because the subject matter is very familiar, but because the full theory would lead us into questions of a depth and complexity too great for an article of this kind.

Both rotations and Lorentz transformations are familiar elements in the edifice of classical physical theory. So are the so-called gauge transformations of the electromagnetic potentials, which are the subject matter of § 4.1. In elementary theory the symmetry principle involved here hardly reveals its power but in modern field theory it becomes an effective tool, ensuring as it does the conservation of electric charge. We have, nevertheless, seen fit to restrict ourselves to a description of selected points of importance for recent theoretical developments and have avoided the use of field theory concepts as much as possible. § 4.2 deals briefly

with other conservation laws of modern theory which are related in form to electric charge conservation.

The symmetry operations mentioned hitherto are described mathematically by *continuous* groups of transformations. There exist in addition symmetry operations of a discrete kind, and § 5 is devoted to these. Foremost among them is reflectional symmetry and the notion of spatial parity associated with it. Until very recent times the invariance of physical laws against reflection—loosely the equivalence of left-handed and right-handed descriptions of physical processes—was accepted as a useful fact, not as intuitively compelling as for example the fact of rotational symmetry, but nevertheless universally true.

The interest in this topic has, however, recently increased greatly with the discovery that in fact physical laws do *not* universally possess reflectional symmetry. What would previously have been treated as one of the absolute symmetries of nature must now be recognized as a relationship that is only approximately valid and it is the *violation* of reflectional symmetry that promises to provide, indeed has already provided, new insight into physical reality. The discussion of reflections (§ 5.1) and of two other discrete symmetry operations closely related to them, namely charge conjugation (§ 5.2) and time reversal (§ 5.3) is the subject of the next chapter, where we have striven to present a rather detailed discussion of current views in a still rather new field of study. This chapter is inevitably rather more technical than most of the preceding sections, but again we have avoided the use of the full machinery of field theory. We have in particular treated in considerable detail the theory of time reversal, of which, as far as we are aware, no comprehensive survey exists at this level.

The recent results on the approximate nature of reflectional invariance have given an added impetus to the study of further symmetries in the laws of nature, which are peculiar to the Quantum Theory of Elementary Particles, and which have always been known to be of approximate character only. The oldest of these is connected with the concept of isobaric spin (this term is now preferred for what was originally rather inappropriately called ' isotopic spin '), to which § 6 is devoted. We discuss this notion in some detail and hope to have brought out how a concept originally introduced as nothing more than a formal mathematical trick, convenient for the description of protons and neutrons in nuclei, has gradually shown itself to possess such a degree of validity and of usefulness that one cannot help feeling today that it points to a symmetry in nature (albeit approximate) of significance comparable to the better understood symmetries discussed earlier.

§ 7 continues this discussion, applying the isobaric spin concept to the whole range of new particles† now known to exist and introducing related notions, such as strangeness, whose usefulness is evidence of yet further symmetries in the structure of natural laws.

Here we are no longer dealing with fully understood and completely proven theories and our report presents a survey of many attempts to put into some order an increasing number of observations in high energy physics. None of the invariance principles suggested can claim to have any deep justification, but it is

† More precisely to the baryons (nucleons, Λ, Σ and Ξ particles) and the mesons (π and K) excluding the μ-meson which together with the electron and the neutrino is thought of as forming the ' lepton ' group, for which the concept of isobaric spin appears to have little meaning.

by now rather convincingly evident that they are expressions of some as yet ill-understood basic relationships. With nothing better than analogies with the rotation groups to guide one, and with no apparent physical meaning for the 'space' in which the quasi-rotations are performed, this formal approach is nevertheless until now the only fruitful one in bringing order into what is observed.

Inevitably this chapter is an interim report, describing as it does a variety of competing classification schemes, none of which can claim to be established in any fundamental sense. However, the main principles of classification have already emerged very clearly and we may not have to wait too long before a deeper understanding of their meaning is found.

The reader will, no doubt, be conscious of a pronounced lack of uniformity in the treatment of the various topics of this article. We have divided the task among the three of us, according to our respective interests and knowledge. We feel, however, that the different treatments correspond in a large measure to differences in the subject matter and that there would have been more loss than gain if we had attempted to achieve stylistic uniformity where the subject matter works against it.

§ 2. Invariance Principles and Conservation Laws

It is a well-known feature of both classical and quantum mechanics that there is an intimate relation between constants of the motion and the corresponding invariance properties of the dynamical system. For example, the invariance of a system under translations in a given direction implies, and is implied by, the conservation of the component of the momentum in that direction.

We can see this for a quantum mechanical system in the following simple manner: If the measuring apparatus is displaced a distance δx in the given direction then we expect that the state vectors and observables of the displaced system will be connected with the state vectors and observables of the original system by a unitary transformation U:

$$|\Psi_d\rangle = U|\Psi\rangle,$$
$$O_d = UOU^{-1}. \qquad \ldots\ldots(2.1)$$

U is a unitary operator satisfying

$$U^\dagger U = UU^\dagger = 1, \qquad \ldots\ldots(2.2)$$

where $U^\dagger$ is the Hermitian conjugate of U, and the suffix d distinguishes the displaced from the original system. The existence of the transformation (2.1) arises from the fact that it preserves the form of any symbolic relation between state vectors and observables. These relations express the physical content of the theory and this is clearly unaltered by a displacement of the measuring apparatus. If the displacement δx is infinitesimally small we may write

$$U = 1 - i\delta x\, P/\hbar \qquad \ldots\ldots(2.3)$$

where the unitarity condition (2.2) will imply that

$$P^\dagger = P; \qquad \ldots\ldots(2.4)$$

that is that P is Hermitian. The resulting change in any observable O is given by

$$\delta_x O = (1 - i\delta x\, P/\hbar) O (1 + i\delta x\, P/\hbar) - O$$
$$= -\delta x [O, P]/i\hbar \qquad \ldots\ldots(2.5)$$

square brackets representing the commutator, $OP - PO$. We may use (2.5) to identify P with the momentum in the x direction. For example, if we take the distance x as our observable, then the new value of the observable in the displaced system is $x - \delta x$. Therefore

$$[x, P] = i\hbar \qquad \ldots\ldots(2.6)$$

which is just the correct commutator for x and its conjugate momentum P. If the system is invariant for translations in the x-direction then the Hamiltonian must remain unchanged under such a translation:

$$\delta_x H = 0. \qquad \ldots\ldots(2.7)$$

This yields

$$[H, P] = 0 \qquad \ldots\ldots(2.8)$$

which is just the quantum condition that P should be a constant of the motion. Conversely, if (2.8) holds we deduce (2.7) which says that the system is translationally invariant.

This type of proof may be generalized for any quantum mechanical observable (or for a classical dynamical variable). If an infinitesimal unitary (canonical) transformation is generated by a function $\epsilon F(q, p)$, where ϵ is small and $F(q, p)$ is a given function of the coordinates and momenta, then the resulting change in the Hamiltonian is given by

$$\delta H = \epsilon \{H, F\}. \qquad \ldots\ldots(2.9)$$

The curly brackets denote $-i/\hbar$ times the commutator for a quantum mechanical system, or the Poisson bracket for a classical system. If in fact this infinitesimal transformation leaves H unchanged then

$$\{H, F\} = 0. \qquad \ldots\ldots(2.10)$$

However since $F(q, p)$ is assumed not to be an explicit function of the time, its total time derivative is given by

$$\frac{dF}{dt} = \{F, H\} = 0. \qquad \ldots\ldots(2.11)$$

We see therefore that those functions which generate infinitesimal transformations leaving H unchanged are all constants of the motion, and conversely. The group of transformations possessing this property will be determined by the symmetry properties of H, which are in turn determined by the symmetry properties of the dynamical system. Thus each symmetry property determines a constant of the motion, namely the generator of the corresponding infinitesimal transformation. This statement needs a little clarification for the quantum case as not all such constants of the motion may be simultaneously observable. The conservation law then states that their probability distributions are independent of time.

An alternative and elegant way of obtaining these relations is in terms of a variational principle. This is particularly useful for the quantum field theoretical

case. The variational principle for quantum theory has been given by Schwinger (1951):

$$\delta\langle\zeta'',\sigma_2|\,\zeta',\sigma_1\rangle = \langle\zeta'',\sigma_2|\,\delta\int_{\sigma_1}^{\sigma_2} L(x)\,d^4x\,|\,\zeta',\sigma_1\rangle. \qquad \ldots\ldots(2.12)$$

The state vectors are eigenstates of the observables ζ on the spacelike surfaces σ_1 and σ_2; and $L(x)$ is the Lagrangian density, a functional of the field operators $\phi(x)$. We may write

$$\delta\int_{\sigma_1}^{\sigma_2} L(x)\,d^4x = F_\delta(\sigma_2) - F_\delta(\sigma_1) + I_\delta, \qquad \ldots\ldots(2.13)$$

where

$$F_\delta(\sigma) = \int_\sigma d\sigma_\mu \frac{\partial L}{\partial\phi_\mu}\,\delta\phi, \qquad \ldots\ldots(2.14)$$

$$I_\delta = \int_{\sigma_1}^{\sigma_2}\left[\frac{\partial L}{\partial\phi} - \frac{\partial}{\partial x_\mu}\left(\frac{\partial L}{\partial\phi_\mu}\right)\right]\delta\phi\,d^4x. \qquad \ldots\ldots(2.15)$$

Since the operators obey the equations of motion

$$I_\delta = 0, \qquad \ldots\ldots(2.16)$$

and so

$$\delta\langle\zeta'',\sigma_2|\,\zeta',\sigma_1\rangle = \langle\zeta'',\sigma_2|\,F_\delta(\sigma_2) - F_\delta(\sigma_1)|\,\zeta',\sigma_1\rangle. \qquad \ldots\ldots(2.17)$$

If the infinitesimal variation δ corresponds to a change that leaves the system unaltered because it possesses the appropriate symmetry property, then the value of $\langle\zeta'',\sigma_2|\,\zeta',\sigma_1\rangle$ must be unchanged for all vectors $|\,\zeta',\sigma_1\rangle$ and $|\,\zeta'',\sigma_2\rangle$. From this we conclude that for this variation

$$F_\delta(\sigma_2) = F_\delta(\sigma_1). \qquad \ldots\ldots(2.18)$$

Clearly equation (2.18) yields a conservation law. In this derivation we use both the invariance property and the equations of motion. The same is true in the earlier discussion as (2.11) is a consequence of the Hamiltonian equations of motion.

This relation between invariance properties and conservation laws is of great importance. The question is often asked how we can be sure that the special theory of relativity with its discussion of space–time structure in terms of such macroscopic ideas as clocks and measuring rods, holds good for microscopic phenomena. The experimental test of this is to be found in the observation that invariance under the inhomogeneous Lorentz group implies the conservation of energy, momentum, and angular momentum.† Thus the measure of our confidence in the truth of these conservation laws for microscopic phenomena is a measure of the invariance of the phenomena under the transformations of the inhomogeneous Lorentz group.

The approach which we have taken to invariance principles and conservation laws, which adds much to our understanding of them, is obviously adapted only to those invariance principles which are associated with a continuous group of transformations, such as displacements. The impossibility of a strictly analogous treatment for discrete transformations such as space inversion, which mathematically is due to the absence of a corresponding infinitesimal operation, is also closely

† The precise connection between the homogeneous Lorentz group and angular momentum is somewhat complicated. It is discussed below in §3.1.

related to the absence of any classical analogue of the conserved quantity, namely parity in the case of space inversion.

Recently, there has been some interest in a type of inexact invariance principle, in which the form of the Lagrangian or the Hamiltonian remains invariant but the values of the c-number parameters of the theory are altered. We have thought it worth while to include a clear discussion of the underlying principles involved, in view of their past and probable future usefulness, but since we make no explicit use of the method in the body of the article we have relegated this account to an Appendix.

§ 3. Rotations and Lorentz Transformations

3.1. *Rotations in Three Dimensions*

Perhaps the most important symmetry in quantum theory is symmetry with respect to rotations about a fixed point in three-dimensional space. A full discussion of the significance of this symmetry would be both inappropriate and unnecessary. In one form or another every textbook on quantum mechanics gives the theory of angular momentum which is nothing other than a statement in physical language of the theory of the representations of the group of rotations. (Monographs on the quantum theory of angular momentum have been published recently by Feenberg and Pake (1953) and, at a more advanced level, by Edmonds (1957) and Rose (1957).) Indeed, the properties of angular momentum and its conservation can be deduced for a system simply from the assumption that there is no preferred direction in space.

The general relation between infinitesimal symmetry operations and quantum operators stated in § 2 finds its particular application to rotations in terms of the equations†

$$R_x = \left(I + \frac{i}{\hbar}\delta\phi M_x\right); \quad R_y = \left(I + \frac{i}{\hbar}\delta\phi M_y\right); \quad R_z = \left(I + \frac{i}{\hbar}\delta\phi M_z\right), \quad \ldots\ldots(3.1)$$

where $\delta\phi$ is an infinitesimal angle. The kinematical relation between the rotations R_x, R_y and R_z about three perpendicular axes immediately gives one the familiar commutation rules between the corresponding components of angular momentum M_x, M_y, M_z, which can be written briefly as

$$\mathbf{M} \wedge \mathbf{M} = i\hbar\mathbf{M} \qquad \ldots\ldots(3.2)$$

or, in a form useful for future reference,

$$M_{ij}M_{kl} - M_{kl}M_{ij} = i\hbar[\delta_{ik}M_{jl} + \delta_{jl}M_{ik} - \delta_{il}M_{jk} - \delta_{jk}M_{il}]. \qquad \ldots\ldots(3.3)$$

Equation (3.3) is valid in a general Euclidean space, where i and j run from 1 to n and $M_{ij} = -M_{ji}$. In three dimensions $M_{23} = M_x$, $M_{31} = M_y$, $M_{12} = M_z$.

The main results of the theory for $n = 3$ are of course that the eigenvalues of $\mathbf{M}^2$ compatible with (3.2) or (3.3) are of the form $\hbar^2 j(j+1)$ where j is either an integer or half an odd integer and that, given the value of $\mathbf{M}^2$, any one component of $\mathbf{M}$ is capable of the eigenvalues $-\hbar j, -\hbar(j-1), \ldots, \hbar j$. A set of eigenvectors of,

† In § 2 the discussion was in terms of a change of coordinate system or measuring apparatus; here we adopt the complementary view that the *system* is displaced. This is the reason why in § 2 the transformation involves $-\delta x$ while here we have $+\delta\phi$.

say M_s, one for each of these eigenvalues and all belonging to the same eigenvalue of $\mathbf{M}^2$, then forms an irreducible representation of the rotation group. Equations (3.1) tell us how these eigenvectors transform among themselves on rotation.

Further developments, such as the so-called vector addition rule for angular momenta need not be elaborated here ; we shall confine ourselves to an exceedingly simple illustrative example which will bring out some points of relevance to other topics discussed in this article.

Consider the three functions of position $xf(r)$, $yf(r)$, $zf(r)$. Taken together as the vector field $\mathbf{r}f(r)$ these functions evidently describe a completely spherically symmetric situation. Physically this is evident, but mathematically it comes about as follows : If the system described by this vector field is thought to be rotated in space by a small angle, then (i) the vector $\mathbf{r}f(r)$ must be subjected to a linear transformation of its components, and (ii) since it depends on the variables x, y, z, its functional dependence must be re-expressed in terms of rotated variables x', y', z'.

(i) The infinitesimal rotation of the components of a vector a_x, a_y, a_z (about, say, the z axis) takes the form

$$a_x' = a_x - \delta\phi a_y, \quad a_y' = a_y + \delta\phi a_x, \quad a_z' = a_z, \qquad \ldots\ldots(3.4)$$

and this relation can be re-written as

$$\mathbf{a}' = \left(I + \frac{i}{\hbar}\delta\phi M_z^{(s)}\right)\mathbf{a} \qquad \ldots\ldots(3.5)$$

where
$$M_z^{(s)} = \hbar\begin{pmatrix} 0 & i & 0 \\ -i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

$M_x^{(s)}$ and $M_y^{(s)}$ can be defined similarly : together with $M_z^{(s)}$ they satisfy (3.2) and represent the so-called spin angular momentum matrices of the field. In our example

$$\mathbf{M}^{(s)2} = \hbar^2 s(s+1) = 2\hbar^2, \qquad \ldots\ldots(3.6)$$

so that we are dealing with the case $s = 1$.

(ii) The rotation affects the independent variables x, y and z in such a way that for an arbitrary function $F(x, y, z)$

$$F(x', y', z') = \left(1 + \frac{i}{\hbar}\delta\phi M_z^{(l)}\right)F(x, y, z) \qquad \ldots\ldots(3.7)$$

where
$$\mathbf{M}^{(l)} = -i\hbar\mathbf{r}\wedge\nabla. \qquad \ldots\ldots(3.8)$$

It is easy to verify that if (3.8) is applied to the functions $xf(r)$, etc., the result is $\mathbf{M}^{(l)2} = 2\hbar^2$, so that we also have $l = 1$ in this case. The invariance of $\mathbf{r}f(r)$ in rotations is now seen to result from the fact that

$$(\mathbf{M}^{(s)} + \mathbf{M}^{(l)})\mathbf{r}f(r) = 0. \qquad \ldots\ldots(3.9)$$

We have here an instance of orbital momentum $l = 1$, and spin momentum $s = 1$, combining to give *total* momentum $j = 0$, i.e. what is described in spectroscopy as a $^3P_0$ state.

The notion of orbital momentum can be readily generalized to a system of more than one particle (more than one set of positional coordinates x, y, z), but it is well known that this extension adds no new results concerning sets of possible angular momentum eigenvalues. In mathematical terms the representations of the rotation group occurring for the orbital angular momentum of the many-particle system are the same as those already obtainable for a single particle: they are all representations having integral values of L, the quantum number for the combined orbital momentum of several particles.†

The half-integral representations of the rotation group do not occur for orbital momentum, but only as spin momenta.‡ This is because it proves impossible to construct functions of the coordinates x, y, z ('wave functions') which have all the necessary properties.§ The half-integral representations are double-valued, so that a rotation through 2π about any axis causes a reversal of sign. For instance, for total spin $\frac{1}{2}$ the infinitesimal rotation about the z axis is usually written as

$$R_z u = (I + \tfrac{1}{2} i \delta\phi \sigma_z) u, \qquad \ldots\ldots(3.10)$$

and if the operator σ_z is represented by the usual Pauli matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

it can be shown by integration that the rotation by a finite angle ϕ takes the form

$$R_z(\phi)\, u = \begin{pmatrix} e^{i\phi/2} \\ e^{-i\phi/2} \end{pmatrix} u. \qquad \ldots\ldots(3.11)$$

Evidently if $\phi = 2\pi$ both components of u change sign.

3.2. *Rotations in Four Dimensions*

Rotation groups in space of more than three dimensions have already been briefly referred to in the footnotes. They do not occur in any physical theories in which the 'space' is actual physical space. (It should be stressed that we are at present explicitly excluding the four-dimensional space–time continuum of special relativity.) Hitherto physical applications of the groups with $n > 4$ have not been numerous and we shall discuss only the case $n = 4$. The relations (3) now involve six quantities M_{ij}. If we introduce the new quantities

$$\begin{aligned} \hbar A_{23} &= \tfrac{1}{2}(M_{23} + M_{14}), & \hbar B_{23} &= \tfrac{1}{2}(M_{23} - M_{14}), \\ \hbar A_{31} &= \tfrac{1}{2}(M_{31} + M_{24}), & \hbar B_{31} &= \tfrac{1}{2}(M_{31} - M_{24}), \\ \hbar A_{12} &= \tfrac{1}{2}(M_{12} + M_{34}), & \hbar B_{12} &= \tfrac{1}{2}(M_{12} - M_{34}), \end{aligned} \qquad \ldots\ldots(3.12)$$

† The situation is different for spaces of higher dimensionality (see below); the single particle orbital angular momentum operator does not exhaust all possible representations of orbital angular momentum operators.

‡ It is, on the other hand, worth noting that for groups of rotations in any number of dimensions, all representations are found among those possible for particle spins in the familiar sense; this is equivalent to saying that all irreducible representations of rotation groups are finite dimensional.

§ It is possible to define functions $Y_l^m(\theta, \phi)$, analogous to the usual surface spherical harmonics, for half-integral values of l and m, but these prove unsuitable as angular momentum wave functions because of difficulties concerned with finiteness.

it can be seen directly that all the A_{ij} commute with all the B_{ij} and that each set separately satisfies the same commutation rules as the $M_{ij}\hbar$ for $n = 3$. Since the whole of the representation theory for $n = 3$ follows from the commutation rule (3.2) or (3.3), we can take over the results of that case for the A_{ij} and the B_{ij} separately. We can thus immediately write down the whole scheme of irreducible representations for $n = 4$.

If we put $$A^2 = \sum_{i<j} A_{ij}^2, \quad B^2 = \sum_{i<j} B_{ij}^2 \qquad \ldots\ldots(3.13)$$

the eigenvalues of these quantities will be respectively $j_1(j_1+1)$ and $j_2(j_2+1)$ where both j_1 and j_2 may be either integers or half-odd integers and can take on their various values separately and independently. For a given pair of values (j_1, j_2) a component, A_{12} say, can have any one of the eigenvalues $-j_1, -j_1+1, \ldots, j_1$, and B_{12}, independently, any one of the eigenvalues $-j_2, -j_2+1, \ldots, j_2$. The 'total angular momentum'

$$\mathbf{M}^2 = 2\hbar^2(\mathbf{A}^2 + \mathbf{B}^2) \qquad \ldots\ldots(3.14)$$

thus has the eigenvalues $2\hbar^2[j_1(j_1+1)+j_2(j_2+1)]$ and the quantity

$$\mathbf{N}^2 \equiv \hbar^2(\mathbf{A}^2 - \mathbf{B}^2) = M_{23}M_{14} + M_{31}M_{24} + M_{12}M_{34} \qquad \ldots\ldots(3.15)$$

the eigenvalues $\hbar^2[j_1(j_1+1) - j_2(j_2+1)]$. The matrices of an irreducible representation (j_1, j_2) are $(2j_1+1)(2j_2+1)$-rowed. We shall list the simplest of these representations.

(1) $j_1 = 0, j_2 = 0$. The obvious invariant or scalar representation.

(2) $j_1 = \frac{1}{2}, j_2 = 0$ and $j_1 = 0, j_2 = \frac{1}{2}$. Both these representations are two-dimensional and double-valued. They are the analogues of the spinor representation in three dimensions and evidently very closely related to it.

(3) $j_1 = \frac{1}{2}, j_2 = \frac{1}{2}$. This representation is four-dimensional and single-valued and relates to vectors in four dimensions. In order to see that the representation is single valued it is necessary to return to the operators M_{ij}, which refer directly to rotations in the four-dimensional space, rather than to stay with the $\mathbf{A}$ and $\mathbf{B}$ operators. For example, a rotation through 2π about the 34 plane is a *simultaneous* rotation of 2π about the third axes of *both* the $\mathbf{A}$ and $\mathbf{B}$ 'spaces', and the sign changes, due to each separately, cancel each other out.

(4) $j_1 = 1, j_2 = 0$ and $j_1 = 0, j_2 = 1$. These two three-dimensional representations are single-valued and at first sight it may appear strange to have a set of three quantities of a tensor nature in four-dimensional space. The explanation is that given an antisymmetric tensor F_{ij} in four dimensions (a six-component quantity) we can form the combinations

$$(F_{23}+F_{14}, F_{31}+F_{24}, F_{12}+F_{34}) \quad \text{and} \quad (F_{23}-F_{14}, F_{31}-F_{24}, F_{12}-F_{34})$$

and these sets of quantities transform separately in all four-dimensional rotations.† An example of such quantities were the A_{ij} and B_{ij} used above.

It is of interest to note that the study of single-particle orbital momentum in four dimensions (four-dimensional spherical harmonics) gives us only those representations for which $j_1 = j_2$; putting $j_1 = j_2 = \frac{1}{2}k$, we find $\mathbf{M}^2 = \hbar^2 k(k+2)$ and $\mathbf{N}^2 = 0$.

† Not however in reflections, which will be discussed in §5.1.

3.3. *Lorentz Transformations*

The group of Lorentz transformations is very closely related to the group of rotations in four dimensions. The Lorentz transformations themselves (which can be taken both as defining the group and as furnishing a particular 4×4 matrix representation of it), are related to rotations in a four-dimensional Euclidean space in the following simple manner: If a (finite) rotation in the 34 plane (or zw plane) is given by

$$z' = z\cos\alpha + w\sin\alpha$$
$$w' = -z\sin\alpha + w\cos\alpha, \qquad \ldots\ldots(3.16)$$

and if the substitution $\alpha\to i\beta$ is made, we get

$$z' = z\cos i\beta + w\sin i\beta = z\cosh\beta + iw\sinh\beta$$
$$w' = -z\sin i\beta + w\cos i\beta = -iz\sinh\beta + w\cosh\beta.$$
$$z' = z\cosh\beta + iw\sinh\beta$$
$$iw' = z\sin\beta + iw\cosh\beta, \qquad \ldots\ldots(3.17)$$

and putting $\quad iw = ct, \quad (iw)' = ct',$

and also $\quad \cosh\beta = (1-v^2/c^2)^{-\frac{1}{2}}$ and $\sinh\beta = (v/c)(1-v^2/c^2)^{-\frac{1}{2}}$

(consistent with $\cosh^2\beta - \sinh^2\beta = 1$), we get

$$z' = (1-v^2/c^2)^{-\frac{1}{2}}[z+vt]$$
$$t' = (1-v^2/c^2)^{-\frac{1}{2}}[t+(v/c^2)z] \qquad \ldots\ldots(3.18)$$

which is the well-known form of a Lorentz transformation. While the original rotation was such as to conserve the value of z^2+w^2 the new transformation conserves $z^2-c^2t^2$. Restating this result in the language of infinitesimal transformations, we can put

$$R_{34} = I + \frac{i}{\hbar}\delta\alpha M_{34} \qquad \ldots\ldots(3.19)$$

and the passage to the Lorentz group is then effected by replacing this by

$$L_{30} = I + \frac{i}{\hbar}\delta\beta M_{30}{}^{L} \qquad \ldots\ldots(3.20)$$

where

$$M_{30}{}^{L} = iM_{34}. \qquad \ldots\ldots(3.21)$$

It follows that if a set of matrices is given which satisfies (3.3) for $n=4$, the replacement

$$M_{i4}\to -iM_{i0}{}^{L}, \quad M_{ij} = M_{ij}{}^{L} \quad (i,j=1,2,3) \qquad \ldots\ldots(3.22)$$

leads directly to a representation of the Lorentz group.

As long as we are dealing with spin transformations for integral angular momentum (for which the column matrices being transformed may be thought of as vectors or tensors in a real space), it is clear that the infinitesimal rotation matrices can be taken to be wholly real, in which case the spin matrices defined by (3.1) are necessarily matrices with pure imaginary elements. By general quantum mechanical principles they must also be Hermitian if they are to represent real physical observables, hence it follows that they are necessarily antisymmetric. Given Hermitian M_{ij} and relation (3.19), the rotation matrices R_{ij} must be unitary (see § 2.2), which fact could also be deduced from the observation

that any vector or tensor in a Euclidean space possesses a positive definite quadratic invariant (for the vector, for example, $a_1^2+a_2^2+a_3^2+a_4^2$).

When one is dealing with half-odd integral representations the unitarity of R_{ij} and the hermiticity of M remain true although it is impossible in general to express the rotations of spinor quantities in terms of real matrices.† In both cases, however, the replacement of α by $i\beta$, which takes one over from the rotation group to the Lorentz group, destroys the unitarity property of the R matrices. (The quadratic invariants of the Lorentz group are not positive definite.) In the case of integral spin representations we can maintain the reality of the R but the M become imaginary and symmetric.‡

This apparently slight formal change has very far-reaching consequences for the mathematical theory underlying our discussion. The so-called 'unitary trick' just employed to produce irreducible non-unitary representations of the Lorentz group starting from known (unitary) representations of the rotation group in four dimensions, works only at the expense of losing all assurance of having obtained *all* irreducible representations of the Lorentz group. Indeed, it works only for the finite or spin representations and for the Lorentz group these by no means exhaust all representations.

The full theory of representations of the Lorentz group is a very much more complicated mathematical problem which we shall not attempt to discuss here. We confine ourselves to pointing out that the problem also corresponds to a much more complex physical situation. When discussing the three-dimensional group we distinguished between spin and orbital angular momentum. In dealing with the latter we had to consider functions of the three coordinates x, y, z. The corresponding discussion for the Lorentz group would evidently involve functions of x, y, z *and the time* t. The study of such representations would thus concern a dynamical system not just at a fixed time t, but at all times, showing that a full knowledge of a representation would imply knowledge of a complete solution of the problem. One is very far from being able to achieve this for systems of physical interest. Unlike the spin representations we have discussed, the representations of greatest interest for quantum mechanical interpretation should be unitary. It is known mathematically that such unitary representations exist.

We see, therefore, that the content of our treatment of the Lorentz group is of rather limited significance. We have been able to show no more than how spinors, vectors and tensors transform. We have also rather lost sight of our general principle that the symmetries of a system are associated closely with physical conservation laws. We know that invariance with respect to spatial

† For spin $\frac{1}{2}$ reality of all spinor components can however be achieved by replacing the original 2×2 rotation matrices $I+\frac{1}{2}\delta\phi i\sigma$ by real 4×4 matrices giving the transformation law as it appears when applied to the column vector

$$\begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{pmatrix} \text{ instead of the original column vector } \begin{pmatrix} u_1+iv_1 \\ u_2+iv_2 \end{pmatrix}.$$

Here $u_1, u_2, \ldots, v_1, v_2$ are real. This is the so-called Majorana representation of a spinor.

‡ We have thought it useful to stress reality condition more than is usually done, in order to fit in with recent ideas on the continuation of the Lorentz group into the complex plane.

rotations gives us conservation of angular momentum. What do 'space–time rotations', i.e. Lorentz transformations, give? What is the physical meaning of those general M_{ij} for which one suffix refers to the time dimension? It is, in fact, not difficult to give a general answer to this question and the answer is rather a simple one, but we shall not derive it in any great generality. Leaving aside all the difficult aspects of the theory, we shall make the result plausible with the aid of a simple example. We know that for a spinless single particle the angular momentum operator is $\mathbf{M} = -i\hbar\mathbf{r}\wedge\nabla$. The behaviour of this quantity as regards spatial rotations only is of course that of an antisymmetric tensor and we should expect that in four-dimensional space–time, it would be related to a six-component antisymmetric tensor. However, a little care is needed in making this extension. One must distinguish between the angular momentum *density* $\mathbf{M}$ and the total angular momentum which will appear in the theory as a quantity of the form

$$\int\psi^{\dagger}(-i\hbar)\mathbf{r}\wedge\nabla\psi\,dV. \qquad \ldots\ldots(3.23)$$

It is in fact the latter quantity that should be generalized. Without entering into discussion of details we can take it that the space–time components of the antisymmetric tensor will have a structure roughly of the form

$$\int\psi^{\dagger}(-i\hbar)\left(ct\nabla+\mathbf{r}\frac{1}{c}\frac{\partial}{\partial t}\right)\psi dV. \qquad \ldots\ldots(3.24)$$

Dividing this expression by mc, where m is the mass of the particle concerned, we see that the first term is of the nature of t/m times the average momentum of the particle. The second term involves $\mathbf{r}$ and represents the negative of a kind of weighted average $\bar{\mathbf{r}}$ of the particle position. Thus if a quantity of the general nature of (3.24) is proved to be equal to a constant, this implies a relation of the type

$$\bar{\mathbf{r}} = \bar{\mathbf{v}}t + \text{const.} \qquad \ldots\ldots(3.25)$$

and we see that the 'conservation law' involved is the statement that the system possesses a suitably definable mass centre whose motion is always uniform. Without making the discussion any more general or precise it is clear from general principles that, in common with classical dynamics, all Lorentz invariant theories will have an integral of this nature and that the statement of its existence is the conservation law involved.

§ 4. Gauge Transformations

4.1. *Gauge Transformations in Electrodynamics*

We now turn to the discussion of an invariance principle of a rather different nature. It involves a group of transformations which has nothing to do with geometry or kinematics, but which nevertheless is well known in classical physics. This principle is called gauge invariance, a name not in the least descriptive of its physical significance.† Gauge invariance states essentially that the potentials of electromagnetic theory are not physically observable quantities and may be subjected to a wide range of transformations without affecting the physical content

† The name 'gradient invariance' used in the Russian literature is much more apt.

of a theory. As is well known, the scalar potential ϕ and the vector potential $\mathbf{A}$ are introduced by the equations

$$\mathbf{B} = \nabla \wedge \mathbf{A}, \quad \mathbf{E} = -\nabla\phi - \dot{\mathbf{A}}/c. \qquad \ldots\ldots(4.1)$$

These relations do not determine ϕ and $\mathbf{A}$ uniquely. If one puts

$$\mathbf{A}' = \mathbf{A} + \nabla\chi, \quad \phi' = \phi - \dot{\chi}/c \qquad \ldots\ldots(4.2)$$

equations (4.1) remain unchanged.

In classical electromagnetic theory these facts are familiar but since it is always explicitly clear that ϕ and $\mathbf{A}$ are auxiliary variables, no great significance is apparent for these gauge transformations. They are, of course, of considerable practical utility; as is well known, χ can be so chosen that the 'Lorentz condition'

$$\nabla . \mathbf{A} + \dot{\phi}/c = 0 \qquad \ldots\ldots(4.3)$$

is satisfied, and if this is done, the field equations of electromagnetism reduce to four separate wave equations, one for ϕ and three for the components of $\mathbf{A}$ separately. This form of the equations still leaves a freedom to transform ϕ and $\mathbf{A}$, but now only with functions χ satisfying the homogeneous wave equation.

The point in classical theory at which gauge transformations obtrude in a non-trivial way is when one attempts to write the equations of motion of a charged particle in an electromagnetic field in Hamiltonian form. The standard expression for such a Hamiltonian is

$$H = \frac{1}{2m}\left(\mathbf{p} - \frac{e}{c}\mathbf{A}(\mathbf{r})\right)^2 + e\phi(\mathbf{r}). \qquad \ldots\ldots(4.4)$$

Here $\mathbf{r}$ is the particle position and $\mathbf{p}$ the momentum conjugate to it in the sense of classical dynamics. Clearly the transformations (4.2) alone do *not* leave the Hamiltonian (4.4) invariant. Does this mean that in the interaction of a particle with an electromagnetic field the effect of two sets of potentials related by (4.2) may be physically different? It does not, and this can be demonstrated by an analysis of the meaning of the Hamiltonian theory and of how the quantities entering it relate to observation. It is easy to show that the 'momentum' $\mathbf{p}$ of Hamiltonian theory is not what is intuitively and physically defined as momentum—mass times velocity. Instead $m\dot{\mathbf{r}} = \mathbf{p} - e\mathbf{A}/c$. A gauge transformation performed on the electromagnetic potentials must thus be coupled with a transformation of the dynamical variables $\mathbf{p}$ (and H) of the particle, in a way permitted by dynamical transformation theory, and like ϕ and $\mathbf{A}$, H and $\mathbf{p}$ are not themselves observable quantities.

This feature of the situation carries over, *mutatis mutandis*, into quantum mechanics where the Hamiltonian (4.4) is used to define the Schrödinger equation of a particle

$$\left[\frac{1}{2m}\left(\frac{\hbar}{i}\nabla - \frac{e}{c}\mathbf{A}(\mathbf{r})\right)^2 + e\phi(\mathbf{r}) + \frac{\hbar}{i}\frac{\partial}{\partial t}\right]\psi = 0. \qquad \ldots\ldots(4.5)$$

Here it is easy to see explicitly how the physical content of (4.5) may be made independent of the transformation (4.2). It is a well-known fact in wave mechanics that the phase of a Schrödinger wave function is of no direct physical significance,

i.e. that if *all* ψ's that may be linearly combined suffer the same transformation

$$\psi' = e^{i\alpha}\psi \qquad \ldots\ldots(4.6)$$

the physical content of the theory does not change. It is easy to verify that if in (4.6) we put

$$\alpha = e\chi/c \qquad \ldots\ldots(4.7)$$

the Schrödinger equation remains invariant under (4.2) and (4.6) and that all equations involving physically observable quantities (in which ψ-functions necessarily enter multiplied by their complex conjugates, and in which $\mathbf{E}$ and $\mathbf{B}$, not ϕ and $\mathbf{A}$, must be involved) are in fact entirely unaffected. As in classical theory, the 'kinetic momentum' operator $\mathbf{p}-e\mathbf{A}/c$, not $\mathbf{p}$, is the physically observable quantity.

The existence of gauge transformations in quantum theory, and their form, is thus established. They consist of the transformations (4.2) on the electromagnetic potentials and the transformation (4.6) and (4.7) on the ψ-functions of particles bearing the electric charge e. (In quantum field theory (4.6) and (4.7) will apply to wave *operators*.) Pauli has called (4.6) the gauge transformation 'of the first kind' and (4.2) the gauge transformation 'of the second kind'. The interesting question now is: what is the conservation law ensured by the existence of the gauge group? The answer is the natural one—conservation of electric charge; we must now show how this comes about. Unfortunately, it is not possible to give a formal demonstration of this statement without going into some—not very difficult—details involving field theory techniques. In ordinary particle dynamics electric charge enters as a constant from the start, not as a dynamic variable; consequently, the apparatus of particle mechanics is quite inadequate for the present discussion and we must study situations in which particle creation and annihilation occurs. However, of the formulae of field theory, we need only use the feature that if a wave operator ψ is associated with the annihilation of some charged particle its Hermitian conjugate, denoted by ψ^*, describes creation of a particle of the same kind. In a gauge transformation ψ operators will be transformed by (4.6), while their Hermitian conjugates will be multiplied by the complex conjugate factor $e^{-i\alpha}$. We see, therefore, that the postulate that a field theoretical Hamiltonian be invariant under (4.6)† is equivalent to the requirement that any term in it contains as many factors creating charged particles as there are factors annihilating them, which is just charge conservation.

Thus the manner in which gauge invariance leads to charge conservation appears at first sight to be quite trivial. Without entering rather more deeply into formal developments than the framework of this article will allow, it is a little difficult to make clear that in fact a fundamental and powerful principle is involved. That this is so may perhaps be shown by a series of not very closely connected remarks.

(i) The techniques briefly described in §2 can be effectively applied in the case that the variations denoted there by $\delta\phi$ are those of the gauge transformation. By such methods one can find for any theory which is of so-called local character (i.e. is one in which statements about physical conditions at a single point in

† If α is taken to be constant no accompanying transformation of the potential occurs.

space-time are meaningful), not only the operator of total charge but also those of charge *density* ρ and current *density* $\mathbf{j}$, and gauge invariance ensures that the latter satisfy an equation of continuity

$$\nabla . \mathbf{j} + \dot{\rho} = 0. \qquad \ldots\ldots(4.8)$$

(ii) In writing down the interaction of charged particles with the electromagnetic field, a quantity e of the dimensions of charge enters the initial equations, as can for instance be seen in (4.4) and (4.5). This constant is, however, in the first place not to be thought of as the physically observed charge, but as the so-called 'bare charge'. The interactions one envisages are such that their presence will modify the charge attached to a material particle. For instance, in meson theory a proton must be thought of as accompanied by a cloud of charged mesons surrounding it, and it is the charge of the combined system that is experimentally observed. Since different types of particles are subject to different interactions, it would seem mysterious at first sight how all observed particles carry the same charge. It can be shown, however, by a direct application of the principle of gauge invariance, that all conceivable 'renormalizations' of charge which might affect different particles in different ways must necessarily vanish. Gauge invariance thus contributes to our understanding of particle structure in a very fundamental way.

(iii) The proof of charge conservation requires rather less than the full principle of gauge invariance. We only needed to use transformations with constant α which need no accompanying transformation (4.2) of the electromagnetic field. It is a tempting speculation to look upon the need to extend gauge transformations to variable α as the *raison d'être* of the electromagnetic field. The argument runs as follows: The physical meaning of a wave function or wave operator ψ is such that quite general changes of phase (gauge transformations of the first kind) should not affect physically observable quantities. However, equations such as the Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + \frac{\hbar}{i}\frac{\partial}{\partial t}\right]\psi = 0 \qquad \ldots\ldots(4.9)$$

are not invariant under such general transformations. One can achieve general invariance by passing from (4.9) to (4.5), i.e. by introducing the electromagnetic field and its gauge transformation (4.2).

It is a matter of philosophy rather than physics to say then that this is the reason why electromagnetic fields exist, but if one accepts the philosophy it is only reasonable to expect it to work in other cases. In current theories there are at least two other quantities besides electric charge, namely baryon and lepton numbers, which appear to be conserved and which, in a local transformation, will be associated with a density and a current density related by an equation such as (4.8).

These continuity equations can again be taken to be consequences of 'gauge transformations' with a constant α, but, despite many attempts, no convincing theories have been produced as yet in which these transformations are extended to general α and the new field then required by the theory is associated with something observed in nature.

Stating much the same in a somewhat different way we can remark that whereas for the electromagnetic case equation (4.8) can be viewed as a consequence of Maxwell's equations

$$\nabla \wedge \mathbf{B} - \frac{1}{c}\dot{\mathbf{E}} = \frac{4\pi}{c}\mathbf{j}\,; \quad \nabla . \mathbf{E} = 4\pi\rho, \qquad \ldots\ldots(4.10)$$

in all other cases the continuity equation appears to be valid in its own right, unrelated to any field equations such as (4.10). One cannot help feeling that one faces a situation in which invariance properties have not yet been exploited as successfully as they might be.

4.2. *Conservation of Baryons and Leptons*

We have just seen the connection between gauge transformations of the first kind and conservation of electric charge. We shall briefly summarize the argument. Invariance of the Hamiltonian under a phase change of all the fields representing charged particles, where the phases for different fields are chosen proportional to the charges, leads to a charge and current density which satisfies the continuity equation (4.8) which is the microscopic expression of charge conservation.† If the phase is taken to be constant, there need be no accompanying change of the electromagnetic potentials, and these become irrelevant. Conversely, if charge is conserved, the Hamiltonian can have no matrix elements connecting states of different total charge and must therefore be invariant under a change in the relative phases of such states. Such a change of phases is the sole effect of gauge transformations of the first kind with a constant phase.

The argument works because the charge on a single particle is a constant for each individual particle, independent of its state. Thus conservation of charge is equivalent to a conservation law of numbers of particles: (the number of particles of charge $+e$) $-$ (the number of charge $-e$) = constant. Similarly, whenever we have a conservation law of this type, we must expect to be able to find a corresponding gauge transformation which leaves the Hamiltonian invariant, and corresponding 'charge' density and current density operators which satisfy microscopic continuity equations. At least one such conservation law—conservation of baryons—is well established, and another—conservation of leptons—probably also holds, although experimental evidence is not yet completely conclusive.

If baryons are taken to be proton, neutron, Λ^0, $\Sigma^{\pm}$ and Σ^0, Ξ^0 and Ξ^-, and the corresponding antiparticles (cf. § 5.2) are called antibaryons, then experimentally the number of baryons minus the number of antibaryons is conserved. The corresponding gauge transformation which must leave the Hamiltonian invariant is

$$\psi \to e^{i\alpha}\psi \qquad \ldots\ldots(4.11)$$

with the same constant phase for all baryon fields, while other fields are left unaltered.

If leptons are defined to comprise the (negative) electron, μ^-, and neutrino, then all experimental evidence appears to be consistent with conservation of

† This assumes a local theory. This restriction can be easily removed (cf. the treatment in the preceding section).

leptons. In this case, however, problems arise because of the non-conservation of parity (see § 5) in all processes in which at least two leptons of different types participate. These problems are related to the fact that neutrinos polarized along their momentum can be regarded as different particles from those polarized in the opposite sense (cf. § 5.1), so that it is permissible to assign different lepton 'charges' to different neutrino polarizations. The neutrino emitted along with a negative electron in beta decay appears to be completely circularly polarized in a right-handed sense, and must be given a leptonic charge opposite in sign to that of the negative electron. The antiparticle of the right-handed neutrino is left-handed, and there is no evidence for the existence of any other type of left-circularly polarized neutrino. Thus the e^-, μ^- and ν_L particles must all be given the same leptonic charge, whereas e^+, μ^+, and ν_R all have the opposite charge.

Since the concepts of left and right circular polarization are not invariant under spatial reflections, we must expect oddities in the behaviour of the leptonic charge and current vector under such reflections: this has been discussed, for example, by Touschek (1957). In fact, the part of the charge and current associated with neutrinos is reversed compared with the part associated with other leptons. It is clear from arguments such as those of § 5.5 that such effects should not be surprising when the only way of determining leptonic charges is through processes where parity is not conserved. For more details, the reader is referred to the original literature.

If the weak interactions responsible for most decay processes are ignored, there is one further conservation law of the type we are considering, which has an associated gauge invariance, and that is conservation of strangeness. The notion of strangeness is discussed in § 7.1.

As already indicated, there is no evidence that any of these conservation laws have associated with them a field analogous to the electromagnetic field associated with electric charge conservation.

§ 5. Space Inversion, Charge Conjugation and Time Reversal

Invariance of the Hamiltonian of the system under the continuous groups of transformations considered in the previous sections leads, as we saw, to conservation laws such as conservation of angular momentum which have well-understood classical analogues. There are also, however, three important discrete transformations whose associated conservation laws have no classical parallel. These transformations will be considered now.

In what follows, we choose our units so that $\hbar$ and c are both equal to 1.

5.1. *Space Inversion*

Let us first consider a classical point particle, whose state is fully described by its position given as a function of time. With any state, described by $\mathbf{r}(t)$, we can associate a *space inverted* state, defined to be the state specified by $\mathbf{r}_P(t) = -\mathbf{r}(t)$. The subscript P is here used to denote space inverted quantities. Any polar vector constructed from $\mathbf{r}$ and its time derivatives is reversed in the space inverted state, and in particular this is true for the linear momentum; any axial vector, such as the angular momentum about the origin, remains unaltered.

Now in the classical theory, the substitutions $\mathbf{p} \to -\mathbf{p}$, $\mathbf{r} \to -\mathbf{r}$ leave the form of the Hamiltonian equations of motion unaltered, so that this is a canonical transformation. Hence the original state can be described in terms of new canonical variables which as functions of time are identical with the functions which describe the space inverted state in the old variables. This change of description of the state, equivalent to space inversion in the sense indicated, is the change from a description in terms of a right-handed system of axes to one in terms of a left-handed system, or vice versa.

If this canonical transformation leaves the Hamiltonian unaltered, then it is obvious that if $\mathbf{r}(t)$ is a solution of the equations of motion, then so is $\mathbf{r}_P(t)$, so that the space inverted state of a dynamically possible state is also dynamically possible. This is the significance of invariance under space inversion in the classical theory.

For single particle quantum mechanics, corresponding to any state $|\Psi\rangle$ it is natural to define a space inverted state $|\Psi_P\rangle$ such that all matrix elements of the position and momentum operators are reversed on inverting the states:

$$\langle\Psi_P(2)|\mathbf{r}|\Psi_P(1)\rangle = -\langle\Psi(2)|\mathbf{r}|\Psi(1)\rangle,$$
$$\langle\Psi_P(2)|\mathbf{p}|\Psi_P(1)\rangle = -\langle\Psi(2)|\mathbf{p}|\Psi(1)\rangle. \qquad \ldots\ldots(5.1)$$

It is also possible to define space inverted operators $\mathbf{r}_P$, $\mathbf{p}_P$ so that

$$\langle\Psi_P(2)|\mathbf{r}_P|\Psi_P(1)\rangle = \langle\Psi(2)|\mathbf{r}|\Psi(1)\rangle \qquad \ldots\ldots(5.2)$$

etc., in analogy with equation (2.1). Then obviously

$$\mathbf{r}_P = -\mathbf{r}, \quad \mathbf{p}_P = -\mathbf{p}. \qquad \ldots\ldots(5.3)$$

Clearly, space inversion of the states is equivalent to space inversion of the operators.

Now since the eigenvalues of $\mathbf{r}_P$ and $\mathbf{p}_P$, and the commutation rules between them, are the same as for $\mathbf{r}$ and $\mathbf{p}$, there must be a unitary operator, P say, such that

$$\mathbf{r}_P = P\mathbf{r}P^{-1}, \quad \mathbf{p}_P = P\mathbf{p}P^{-1}, \qquad \ldots\ldots(5.4)$$

and, from (5.2),

$$|\Psi_P\rangle = P|\Psi\rangle. \qquad \ldots\ldots(5.5)$$

The unitary transformation P is the quantum analogue of the canonical transformation of the classical theory already considered.

Equations (5.4) define P only up to a phase which remains arbitrary. However, it is obvious that the physical significance of state vectors must remain unaltered by a double application of space inversion, so that P^2 must be just a phase factor. It is usual to choose

$$P^2 = 1. \qquad \ldots\ldots(5.6)$$

P is now determined apart from an arbitrary sign. The unitary character of P together with (5.6) shows that P is Hermitian, with eigenvalues ± 1. P is called the *parity operator*.†

We confine attention at present to the case of a particle without spin. If $\psi(\mathbf{r})$ is the wave function for the state $|\Psi\rangle$, it is obvious that the wave function for $|\Psi_P\rangle$ is

$$\psi_P(\mathbf{r}) = \pm\psi(-\mathbf{r}). \qquad \ldots\ldots(5.7)$$

† P does not correspond to any classical observable quantity—this is discussed later.

The choice of sign in (5.7), which corresponds to the remaining ambiguity of sign in the definition of P, can be made arbitrarily for each type of particle, but must be the same for all states of the same particle. The sign, once chosen, is called the *intrinsic parity* of the particle. If the state of the system is an eigenstate of the operator P, then $\psi_P(\mathbf{r}) = \pm\psi(+\mathbf{r})$, where the sign is that of the eigenvalue of P and is the *total parity* or just the parity of the state. In this case, $\psi(-\mathbf{r}) = \pm\psi(\mathbf{r})$, and the sign here can be called the *orbital parity*. The total parity of a state is the product of its orbital and intrinsic parities. For example, for a π meson in a state of even angular momentum, $\psi(-\mathbf{r}) = +\psi(\mathbf{r})$, so that the orbital parity is $+1$, while the intrinsic parity is -1. Hence by (5.7) and the symmetry of the wave function, $\psi_P(\mathbf{r}) = -\psi(-\mathbf{r}) = -\psi(\mathbf{r})$, and this shows that the total parity is also -1.

If the Hamiltonian of the system is invariant under space inversion, then $|\Psi_P\rangle$ will be a dynamically possible state provided $|\Psi\rangle$ is. This corresponds to the classical result. In quantum mechanics, however, the invariance property has the further consequence that the Hermitian operator P commutes with the Hamiltonian, and therefore represents a quantity, in principle observable, which is a constant of the motion. This is expressed by saying that parity is conserved. The importance of this conservation law lies in the fact that natural processes are usually most conveniently described in terms of states of definite parity, and indeed bound states of definite energy are almost always also states of definite parity. The existence of this conservation law in quantum mechanics, which has no direct classical analogue, can be traced to the validity of the superposition principle in quantum mechanics, which also has no classical counterpart.

We now extend the parity transformation to quantum field theory, in the first instance for free fields. If $\phi(\mathbf{r}, t)$ is a field describing a spinless particle, and if $|0\rangle$ is the vacuum state, then the state $|U\rangle$ in which there is one particle with wave function $u(\mathbf{r})$ is

$$|U\rangle = \int \phi^\dagger(\mathbf{r})\, u(\mathbf{r})\, d^3\mathbf{r}\, |0\rangle \qquad \ldots\ldots(5.8)$$

and many particle states can be defined similarly.† If it is postulated that

$$|0_P\rangle = |0\rangle \qquad \ldots\ldots(5.9)$$

then the effect (5.7) of the parity transformation on a single particle wave function shows that the space inverted field operator $\phi_P(\mathbf{r}, t)$ is given by

$$\phi_P(\mathbf{r}, t) = \pm\phi(-\mathbf{r}, t) \qquad \ldots\ldots(5.10)$$

where the sign is the intrinsic parity of the particle. If the sign is $+$, then the field ϕ is said to be scalar, and if the sign is $-$, ϕ is pseudoscalar. (Kemmer

† It is important in later work to remember how the vacuum is defined. The field $\phi(\mathbf{r}, t)$ may be divided into positive and negative frequency parts as follows:

$$\phi(\mathbf{r}, t) = \phi_+(\mathbf{r}, t) + \phi_-(\mathbf{r}, t)$$

where for example

$$\phi_+(\mathbf{r}, t) = \frac{1}{2\pi}\int_0^\infty dk_0 \int_{-\infty}^{\infty} dt' \phi(\mathbf{r}, t') \exp\{ik_0(t'-t)\}.$$

Then the vacuum state $|0\rangle$ is defined so that

$$\phi_+(\mathbf{r}, t)\,|0\rangle = \phi_-(\mathbf{r}, t)\,|0\rangle = 0.$$

It follows that in fact only the positive frequency part of ϕ contributes to the state $|U\rangle$ defined by (5.8).

(1938 a) was the first to consider a pseudoscalar field, among other possibilities, i.e. a spin zero particle with odd intrinsic parity.) An argument similar to that given for the single particle case shows that in the field theory case too there must exist a unitary parity operator P such that

$$\phi_P = P\phi P^{-1}, \quad |\Psi_P\rangle = P|\Psi\rangle. \qquad \ldots\ldots(5.11)$$

The space inversion transformation for a theory with interacting fields can then be defined by

$$|\Psi\rangle \rightarrow |\Psi_P\rangle,$$

where $$\langle\Psi_P(2)|\phi_P(\mathbf{r}, t)|\Psi_P(1)\rangle = \langle\Psi(2)|\phi(\mathbf{r}, t)|\Psi(1)\rangle,$$

and $$\phi_P(\mathbf{r}, t) = \pm\phi(-\mathbf{r}, t),$$

in accordance with (5.10).‡

A similar discussion can be given for a particle with spin. In this case, however, equations (5.4) are not sufficient to define the single particle parity operator up to a phase, and must be supplemented by other conditions. For example, the intrinsic spin of the particle, being an angular momentum, should transform in the same way as orbital angular momentum, so that if $\mathbf{s}$ is the spin operator,

$$P\mathbf{s}P^{-1} = \mathbf{s}_P = \mathbf{s}. \qquad \ldots\ldots(5.12)$$

Similarly, the velocity operator, $\boldsymbol{\alpha}$ say, should be reversed by space inversion:

$$P\boldsymbol{\alpha}P^{-1} = \boldsymbol{\alpha}_P = -\boldsymbol{\alpha}. \qquad \ldots\ldots(5.13)$$

In a relativistic theory (5.13) is in general distinct from the second of equations (5.4).

The effect of space inversion on the field operators for particles of spins 0, $\frac{1}{2}$ and 1 is summarized in table 1, with the customary choices of phases for the parity operator. The need to satisfy equations (5.12) and (5.13) generally forces different

Table 1

Spin	Type of Field	Symbol	Parity Transformations
0	scalar pseudoscalar	ϕ	$\phi_P(\mathbf{r}, t) = \pm\phi(-\mathbf{r}, t)$
$\frac{1}{2}$	spinor†	ψ	$\psi_P(\mathbf{r}, t) = \pm\gamma_0\psi(-\mathbf{r}, t)$ or $\psi_P(\mathbf{r}, t) = \pm i\gamma_0\psi(-\mathbf{r}, t)$
1	vector pseudovector	ϕ_μ	$\phi_{kP}(\mathbf{r}, t) = \mp\phi_k(-\mathbf{r}, t) \quad (k = 1, 2, 3)$ $\phi_{0P}(\mathbf{r}, t) = \pm\phi_0(-\mathbf{r}, t).$

† We use Dirac matrices γ_0, γ_1, γ_2, γ_3 such that $\gamma_0^2 = -1$, $\gamma_0^\dagger = -\gamma_0$, $\gamma_j^2 = +1$, $\gamma_j^\dagger = \gamma_j$ $(j = 1, 2, 3)$.

‡ We could, alternatively, have tried to define space inversion for interacting fields directly by a correspondence principle argument similar to that used for free fields, but this would have led to difficulties concerned with questions such as the definition of the vacuum, and the relationship between 'bare' and 'real' particle states. Since, however, in scattering processes it is always supposed that the initial and final states are asymptotically free particle states, it is clear that our procedure is consistent with the correspondence principle for such processes even with interacting fields, and it is reasonable to suppose this will be so also for bound state and other problems.

components of the field to have different intrinsic parities, as is seen in particular for the vector and pseudovector in the table. For the spinors too, the signs of the transformation are opposite for those components corresponding to the eigenvalue $+i$ of γ_0, compared with those belonging to the eigenvalue $-i$. The space inversion transformation properties for field operators with other Lorentz transformation characters are easily obtained. For example, for a general tensor component $T_{\lambda\mu\nu\ldots}(\mathbf{r}, t)$, the transformation is

$$T_{\lambda\mu\nu\ldots}(\mathbf{r}, t) \rightarrow \pm(-)^{n_s} T_{\lambda\mu\nu\ldots}(-\mathbf{r}, t) \qquad \ldots\ldots(5.14)$$

where n_s is the number of space-like indices (i.e. 1, 2 or 3) among $\lambda, \mu, \nu, \ldots$.

Certain reservations have to be made, especially in the case of spin $\frac{1}{2}$ fields, and these will now be considered. In the first place, it is well known that the absolute sign of a spinor field has no physical significance, since it can be reversed by a rotation of the system through 2π (cf. § 3.1), which should leave all physically important quantities unaltered. For this reason, it is customary to relax the condition (5.6) and to allow instead

$$P^2 = \pm 1 \qquad \ldots\ldots(5.6')$$

for spin $\frac{1}{2}$ particles. This accounts for the apparent occurrence of four types of intrinsic parity for such particles, as indicated in the table. The transformation $\psi_P(\mathbf{r}, t) = \pm\gamma_0\psi(-\mathbf{r}, t)$ is usually preferred, for reasons concerned with the charge conjugation transformation, to be discussed later. Among the many papers dealing with the parity transformation of spinors, we wish to mention in particular Yang and Tiomno (1950), Wick, Wightman and Wigner (1952), and Morpurgo and Touschek (1955), where further references will be found.

It is also important to remember that the condition (5.6) is artificially imposed, and that the more general condition is that P^2 should be a phase factor.† This phase factor could be different for different particles: we have already described a special case of this in the preceding paragraph. A more detailed discussion is given by Wick *et al.* (1952). It is even possible to satisfy equations (5.6), (5.12) and (5.13) using different phase factors for different states of the same particle. In particular, the phase factor for states with $\mathbf{s}.\mathbf{p} > 0$ may be taken different from the phase factor for states with $\mathbf{s}.\mathbf{p} < 0$. The free particle Hamiltonian is not in general invariant under such a parity transformation, and for this reason such transformations are usually rejected. For zero mass particles, however, the free Hamiltonian is invariant and the more general transformations must be allowed. For a zero mass spin $\frac{1}{2}$ field (e.g. the neutrino field), the most general parity transformation is

$$\psi_P(\mathbf{r}, t) = \gamma_0 \exp\{i\alpha + i\beta\gamma_5\}\psi(-\mathbf{r}, t) \qquad \ldots\ldots(5.15)$$

† If P^2 is a phase factor in the single particle theory, so that the field operator in the field theory transforms according to

$$\phi_P(\mathbf{r}, t) = e^{i\alpha}\phi(-\mathbf{r}, t),$$

then P^2 in the quantized field theory is not in general even a multiple of the unit operator; in such a case double application of space inversion will change the relative phases of states containing different numbers of particles. This is quite consistent with our correspondence principle approach to the definition of the transformation, since we cannot deduce from the correspondence principle anything about the transformation properties of matrix elements of field operators between states having different numbers of particles. The only exception to this is for a Hermitian field operator, such as the electromagnetic field, which is in principle measurable and may have a definite classical correspondence principle limit.

where α and β are arbitrary phase angles. The general transformation (5.15) has proved important in connection with non-conservation of parity in weak decay processes. In one sense, the ambiguity in the choice of β in (5.15) is a special case of the possibility of choosing different phases for the parity transformation for different particles, for the states of a zero mass particle with $\mathbf{s}.\mathbf{p}>0$ are dynamically independent of the states with $\mathbf{s}.\mathbf{p}<0$, and the two classes of states so defined may be regarded if one wishes as states of two different particles. The full ambiguity of the parity transformation for zero mass particles was first described in detail by Shaw (1955).

Since the assignment of intrinsic parity to particles remains arbitrary, there are many parity transformations which satisfy the conditions we have so far required. The Hamiltonian of a system of interacting fields is said to be invariant under space inversion if there is at least one choice of intrinsic parities for the different particles, or more generally one choice of the phase factors associated with the parity transformations of the different fields, so that the Hamiltonian is left unaltered by the complete parity transformation. Normally there will be more than one possible assignment of intrinsic parities. This implies that the concept of intrinsic parity for a particular particle need not have any sharp meaning. Indeed it is clear from the earlier discussion that this is always so for spin $\frac{1}{2}$ particles. However, it may be possible to give sense to the idea of the *relative parity* of two particles. This will be so if for all possible assignments of intrinsic parities, the product of the parities of the two particles is always the same. Similarly, it may be possible to define the parity of one particle relative to a pair of other particles. A fuller account of the determination of intrinsic parities will be given in § 5.5.

In discussing the invariance of a Hamiltonian under space inversion, it is necessary to allow the full generality of the transformation (5.15) for the neutrino field. However, if a Hamiltonian is invariant under a parity transformation in which the neutrino field transforms according to (5.15), it is always possible by a unitary transformation to change it into an equivalent Hamiltonian which is invariant under a parity transformation in which the neutrino field transforms according to table 1, like the other spin $\frac{1}{2}$ fields (Pursey 1957).

5.2. *Charge Conjugation*

The charge conjugate of a given state of a charged particle moving in an electromagnetic field is that state of an oppositely charged particle moving in an exactly reversed electromagnetic field with exactly the same trajectory as the original particle. Although this transformation leaves unaltered the canonical equations of motion of the system of particle and electromagnetic field, it would not normally be regarded as a canonical transformation because the charge would not normally be regarded as one of the generalized coordinates of the particle. The same difficulty prevents the charge conjugation transformation from being a unitary transformation in single particle quantum mechanics. We shall therefore deal directly with quantum field theory, where charge and current densities are operators which can be changed by a unitary transformation. (It is possible to build an unquantized analogue within the framework of classical field theory.) We shall deal for simplicity mainly with a charged scalar field $\phi(x)$.

The state $|U\rangle$ for a free field which describes a single particle with wave function $u(\mathbf{r})$ is, according to (5.8),

$$|U\rangle = \int \phi^\dagger(\mathbf{r})\, u(\mathbf{r})\, d^3\mathbf{r}\, |0\rangle. \qquad \ldots\ldots(5.8)$$

The total charge of the system is

$$Q = \tfrac{1}{2}ie\int\left(\phi^\dagger\frac{\partial\phi}{\partial t} + \frac{\partial\phi}{\partial t}\phi^\dagger - \frac{\partial\phi^\dagger}{\partial t}\phi - \phi\frac{\partial\phi^\dagger}{\partial t}\right)d^3\mathbf{r} \qquad \ldots\ldots(5.16)$$

and it is possible to verify that (5.8) is in fact a state of charge $+e$.† The state $|U_C\rangle$ charge conjugate to $|U\rangle$ will be a state in which there is a single particle of charge $-e$ and with wave function $u(\mathbf{r})$. It is very easily seen that

$$|U_C\rangle = e^{i\beta}\int\phi(\mathbf{r})\, u(\mathbf{r})\, d^3\mathbf{r}\, |0\rangle, \qquad \ldots\ldots(5.17)$$

where β is a phase angle, satisfies this requirement.

We can now define a charge conjugate field $\phi_C(\mathbf{r}, t)$ so that

$$\langle\Psi_C(2)|\phi_C|\Psi_C(1)\rangle \equiv \langle\Psi(2)|\phi|\Psi(1)\rangle, \qquad \ldots\ldots(5.18)$$

and it is easy to show that a possible solution is‡

$$\phi_C(\mathbf{r}, t) = e^{i\beta}\phi^\dagger(\mathbf{r}, t), \quad \phi_C^\dagger(\mathbf{r}, t) = e^{i\beta}\phi(\mathbf{r}, t). \qquad \ldots\ldots(5.19)$$

The commutation relations of the fields are left unaltered by the interchange of ϕ and $\phi^\dagger$, so that there must be a unitary operator C such that

$$\phi_C = C\phi C^{-1}, \quad |\Psi_C\rangle = C|\Psi\rangle. \qquad \ldots\ldots(5.20)$$

C is the charge conjugation operator. It is a unitary operator which, apart from phase factors, transforms ϕ into $\phi^\dagger$ and vice versa. *It does not replace all operators by their Hermitian conjugates.* For example, the Hermitian conjugate of i times the unit operator is $-i$, but $CiC^{-1} = +i$. Similarly, $Ci\phi C^{-1} = i\phi_C = ie^{-i\beta}\phi^\dagger$ whereas $e^{-i\beta}(i\phi)^\dagger = -ie^{-i\beta}\phi^\dagger$. The particle associated with the field ϕ_C is called the *antiparticle* of that associated with the field ϕ.

From (5.16), we see that

$$Q_C \equiv CQC^{-1} = -Q \qquad \ldots\ldots(5.21)$$

and similarly the current vector changes sign.

The transformation is now extended to the case of interacting fields by carrying over (5.19) and (5.20). (5.21) remains valid, so that the charge conjugate of a state has the opposite charge to the original state.

† The commutation rules for the fields are

$$\left[\frac{\partial\phi^\dagger}{\partial t}(\mathbf{r}, t),\ \phi(\mathbf{r}', t)\right] = -i\delta(\mathbf{r}-\mathbf{r}').$$

It follows that $[Q, \phi^\dagger(\mathbf{r})] = e\phi^\dagger(\mathbf{r})$, and since $Q|0\rangle = 0$, $|U\rangle$ must be an eigenstate of Q with eigenvalue $+e$.

‡ Consider, for example, the matrix element of ϕ between a one-particle state U and the vacuum. We wish to show that

$$\langle 0|\phi^\dagger|U_C\rangle = \langle 0|\phi|U\rangle,$$

i.e.

$$\int d^3\mathbf{r}'\,\langle 0|\phi^\dagger(\mathbf{r})\phi(\mathbf{r}')\, u(\mathbf{r}')|0\rangle = \int d^3\mathbf{r}'\,\langle 0|\phi(\mathbf{r})\phi^\dagger(\mathbf{r}')\, u(\mathbf{r}')|0\rangle.$$

If one uses the definition of the vacuum in the footnote to eqn. (5.8), the required result is seen to follow from

$$[\phi_-^\dagger(\mathbf{r}), \phi_-(\mathbf{r}')] = [\phi_+(\mathbf{r}), \phi_+^\dagger(\mathbf{r}')],$$

which is easily proved for free fields.

For a complex field associated with a neutral particle, it is still possible to define a particle–antiparticle transformation by (5.19) and (5.20). We shall still call this 'charge' conjugation. It is possible for a neutral field to be Hermitian: in this case the only possibilities for the phase factor $e^{i\beta}$ in (5.19) are ± 1. The electromagnetic field is an example of a Hermitian field whose phase factor is -1:

$$A_{\mu C} \equiv CA_\mu C^{-1} = -A_\mu. \qquad \ldots\ldots(5.22)$$

This choice is made to ensure the invariance of electromagnetic interactions under charge conjugation. In particular, it ensures that the electromagnetic field generated by a system of particles is reversed when all the charges are reversed by charge conjugation.

For a charged vector meson (spin 1) field, the definition of charge conjugation is essentially the same as for a scalar field. Some modifications may be needed for a spinor field, because the form of the charge and current operators is rather different. For a spinor field ψ, the charge and current operators are

$$Q = \tfrac{1}{2}e\int(\psi^\dagger\psi - \psi\psi^\dagger)\,d^3\mathbf{r}, \qquad \ldots\ldots(5.23)$$

$$J_k = \tfrac{1}{2}e\int(\psi^\dagger\gamma_0\gamma_k\psi - \psi\gamma_k{}^{\mathrm{T}}\gamma_0{}^{\mathrm{T}}\psi^\dagger)\,d^3\mathbf{r}, \qquad \ldots\ldots(5.24)$$

where $\gamma_k{}^{\mathrm{T}}$, $\gamma_0{}^{\mathrm{T}}$ are the transposes of the matrices γ_k and γ_0. It is possible to choose Dirac matrices whose matrix elements are wholly real: such a representation of the matrices is called a Majorana representation. In this case, since the γ_k are Hermitian and γ_0 is anti-Hermitian, $\gamma_k{}^{\mathrm{T}}\gamma_0{}^{\mathrm{T}} = \gamma_0\gamma_k$. It follows that Q and $\mathbf{J}$ change sign correctly if

$$\psi_C = e^{-i\beta}\psi^\dagger, \quad \psi_C{}^\dagger = e^{i\beta}\psi \qquad \ldots\ldots(5.25)$$

in analogy with the scalar case. If, however, we do not choose a Majorana representation for the Dirac matrices (5.25) must be replaced by

$$\psi_C = e^{-i\beta}\mathscr{C}\psi^\dagger, \quad \psi_C{}^\dagger = e^{i\beta}\mathscr{C}^{\mathrm{T}}\psi \qquad \ldots\ldots(5.26)$$

where $\mathscr{C}$ is a 4×4 unitary matrix with the property

$$\mathscr{C}\gamma_\mu\mathscr{C}^\dagger = \gamma_\mu{}^*. \qquad \ldots\ldots(5.27)$$

Here $\gamma_\mu{}^*$ is the matrix whose matrix elements are the complex conjugates of the corresponding matrix elements of γ_μ.† We shall normally use a Majorana representation in all further discussions.

From our definitions, it is easily seen that C and P commute. To show this, it is essential to remember that C does not replace all operators by their Hermitian conjugates: in particular, it will not affect any complex phase factor associated with P. It was stated in the previous section that for spinor fields the transformation $\psi_P(\mathbf{r}) = \pm\gamma_0\psi(-\mathbf{r})$ is usually preferred to the transformation $\psi_P(\mathbf{r}) = \pm i\gamma_0\psi(-\mathbf{r})$. The reason for this is most clearly seen in the condition for a Majorana field, namely

$$\psi = \psi_C, \qquad \ldots\ldots(5.28)$$

† If we change the representation of the Dirac matrices to $\gamma_\mu' = U\gamma_\mu U^{-1}$ where U is some unitary matrix, then we must change $\mathscr{C}$, not to $U\mathscr{C}U^{-1}$ as we might expect, but to $\mathscr{C}' = U^*\mathscr{C}U^{-1}$. This is mentioned because, when we come to deal with time reversal, we shall find a unitary operator occurring which is defined in a way rather similar to (5.27), and in consequence does not change in the usual way under a unitary transformation of the theory.

or, with the phase factor chosen to be unity,

$$\psi = \psi^{\dagger}. \qquad \ldots\ldots(5.29)$$

If, now, the transformation $\psi_{\mathrm{P}}(\mathbf{r}) = \pm\gamma_0\psi(-\mathbf{r})$ is chosen, then with our definition of charge conjugation conditions (5.28) and (5.29) will not be preserved and in particular we shall obtain

$$\psi_{\mathrm{P}} = \psi_{\mathrm{CP}} = -\psi_{\mathrm{PC}} = -\psi_{\mathrm{P}}{}^{\dagger}.$$

This has physically objectionable consequence (cf. Wightman and Schweber 1955).

For a zero mass neutral particle with spin, there is in general an ambiguity in the definition of particle and antiparticle. This arises because, as mentioned in § 5.1, the states with spin parallel to the momentum may be regarded as belonging to a different particle from those with spin antiparallel to momentum: we can call these respectively right and left particles, and their corresponding antiparticles we can call anti-right and anti-left. Then it can be shown that the anti-right particle has its spin anti-parallel to its momentum, so that is really a left-handed particle, and similarly the anti-left particle is really right-handed. Provided that all four types of state exist, it is possible to form mixed states of left and anti-right particles, and of right and anti-left particles, and to construct field operators corresponding to calling the mixed left and anti-right states, states of a new left particle, and similarly for the other mixtures. This was first discussed for the neutrino by Pauli (1957). In this case, the left and right fields are related to the usual Dirac field ψ by

$$\psi_{\mathrm{L}} = \tfrac{1}{2}(1+\gamma_5)\psi, \quad \psi_{\mathrm{R}} = \tfrac{1}{2}(1-\gamma_5)\psi, \qquad \ldots\ldots(5.30)$$

and Pauli's unitary transformation of the fields may be written

$$\begin{aligned} \begin{pmatrix} \psi_{\mathrm{L}} \\ \psi_{\mathrm{RC}} \end{pmatrix} \to \begin{pmatrix} \psi_{\mathrm{L}}' \\ \psi_{\mathrm{RC}}' \end{pmatrix} &= \begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix} \begin{pmatrix} \psi_{\mathrm{L}} \\ \psi_{\mathrm{RC}} \end{pmatrix}, \\ \begin{pmatrix} \psi_{\mathrm{R}} \\ -\psi_{\mathrm{LC}} \end{pmatrix} \to \begin{pmatrix} \psi_{\mathrm{R}}' \\ -\psi_{\mathrm{LC}}' \end{pmatrix} &= \begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix} \begin{pmatrix} \psi_{\mathrm{R}} \\ -\psi_{\mathrm{LC}} \end{pmatrix}, \end{aligned} \qquad \ldots\ldots(5.31)$$

where

$$|a|^2 + |b|^2 = 1. \qquad \ldots\ldots(5.32)$$

Unless the particle has an interaction which unambiguously distinguishes the two left states and the two right states, it will be impossible to give a clear-cut definition of particle and antiparticle.

If, however, there exists only one left and one right state for any given momentum, there is no ambiguity, but there is nevertheless a choice in the description of the system. This may be seen by imposing the Majorana condition in the form (5.29), and defining charge conjugation by $\psi_{\mathrm{LC}} = \psi_{\mathrm{L}}{}^{\dagger}$. Then it is found that $\psi_{\mathrm{R}} = \psi_{\mathrm{LC}}$. (In a Majorana representation, γ_5 is purely imaginary.) Thus one has the choice in a Majorana theory either of treating the left and right particles as antiparticles of each other, as in the two-state neutrino theory of Lee and Yang (1957), Landau (1957) and Salam (1957), or of treating them as different states of the same particle, which is identical with its antiparticle, as in the original theory (Majorana 1937). This equivalence was first demonstrated by Serpe (1952).

The transformations (5.31) are identical in form to the transformation of a spinor under a rotation in three dimensions (see § 3.1). Thus in a sense the

Pauli transformations are equivalent to a rotation in some abstract three-dimensional space. For a particle with mass and charge, such as the electron, the mass and charge can be regarded as parallel vectors in this space (cf. Kahana and Pursey 1957). Alternatively, it is possible to describe a massive particle by a theory invariant under transformations analogous to (5.31) if the number of independent states of the particle is doubled. This has been done by Gürsey (1958), who showed that in this way the transformations (5.31) could be related to the isobaric spin transformations of the nucleon field (see § 6). This has been applied in the later forms of Heisenberg's theory of fundamental particles (CERN 1958).

From (5.19) and (5.25) it appears that C^2 commutes with all operators and must therefore be a multiple of the unit operator. It is usual to put

$$C^2 = 1, \qquad \ldots\ldots (5.33)$$

which is necessary if we require $C|0\rangle = |0\rangle$. From (5.33), it follows that C is Hermitian as well as unitary, and has eigenvalues ± 1.

If the phases β can be chosen for all particles so that the Hamiltonian is invariant under the operation of charge conjugation, then if $|\Psi\rangle$ is a dynamically possible state so also is $|\Psi_C\rangle$. In addition, as for parity, C is a Hermitian operator (*charge parity*) which may represent an observable, and which is conserved. However, it is obvious that C and the total charge cannot be simultaneously sharp, except in the special case when the total charge is zero. Indeed, in a state of definite charge parity, the sign of the charge will be $\pm$ with equal probabilities. Since there is no known way of preparing a pure quantum state in which the charge is not sharp, the concept of charge parity is less useful than the concept of space parity. However, equation (5.22) shows that a state with n photons has charge parity $(-1)^n$. This has importance in determining selection rules for decay processes involving photons : for example, it can be shown that positronium in the ^1S state must decay into an even number of photons, whereas positronium in the ^3S state must go into an odd number of photons.† Similarly, the π^0 must have charge parity $+1$, since it can decay into two but not into three photons.

† Consider, for example, the state

$$|\Psi\rangle = \sum_{\alpha,\beta} \int u_{\alpha\beta}(\mathbf{r}_1, \sigma_1\,;\, \mathbf{r}_2, \sigma_2)\, \psi_\alpha{}^\dagger(\mathbf{r}_1)\, \psi_{C\beta}{}^\dagger(\mathbf{r}_2)\, d^3\mathbf{r}_1\, d^3\mathbf{r}_2\, |0\rangle,$$

where σ_1 and σ_2 are included in the wave function to denote spin dependence. Then

$$C|\Psi\rangle = \sum_{\alpha,\beta} \int u_{\alpha\beta}(\mathbf{r}_1, \sigma_1\,;\, \mathbf{r}_2, \sigma_2)\, \psi_{C\alpha}{}^\dagger(\mathbf{r}_1)\, \psi_\beta{}^\dagger(\mathbf{r}_2)\, d^3\mathbf{r}_1\, d^3\mathbf{r}_2\, |0\rangle,$$

$$= -\sum_{\alpha,\beta} \int u_{\alpha\beta}(\mathbf{r}_1, \sigma_1\,;\, \mathbf{r}_2, \sigma_2)\, \psi_\beta{}^\dagger(\mathbf{r}_2)\, \psi_{C\alpha}{}^\dagger(\mathbf{r}_1)\, d^3\mathbf{r}_1\, d^3\mathbf{r}_2\, |0\rangle,$$

(since $\psi^\dagger$ and $\psi_C{}^\dagger$ anticommute)

$$= -\sum_{\alpha,\beta} \int u_{\alpha\beta}(\mathbf{r}_2, \sigma_2\,;\, \mathbf{r}_1, \sigma_1)\, \psi_\alpha{}^\dagger(\mathbf{r}_1)\, \psi_{C\beta}{}^\dagger(\mathbf{r}_2)\, d^3\mathbf{r}_1\, d^3\mathbf{r}_2\, |0\rangle,$$

on interchanging dummy integration variables and dummy suffixes in the sum over components of u and of the field operators ψ. Now, in an S-state, u is symmetric on interchange of the space co-ordinates of the two particles (this is not quite the same as the parity operation since the question of intrinsic parities is by-passed). u is symmetric or antisymmetric under interchange of spin variables if the state is a triplet or a singlet respectively. Hence

$$C|\Psi\rangle = \pm\,|\Psi\rangle$$

according as $|\Psi\rangle$ is ^1S or ^3S. Of course the choice of state vector is oversimplified in that additional virtual pairs are ignored. A more detailed treatment is given by Jauch and Rohrlich (1955).

If a neutral particle is distinct from its antiparticle, then there is the possibility of states of definite charge parity playing an important role. For example, Gell-Mann and Pais (1955 b) suggested that there might be differences in the decay of the states of the neutral K meson with opposite charge parity. This will be discussed in § 7.1.

It is now believed that neither space parity nor charge parity are conserved in weak interactions, but that the product CP does represent a conserved quantity.†

5.3. *Time Reversal*

In classical physics, the instantaneous state of a particle is regarded as defined when the instantaneous position and velocity are given. We can define a corresponding time reversed state as one in which the position is unaltered but the velocity reversed: if $\mathbf{r}, \mathbf{v}$ refer to the original state, the corresponding quantities for the time reversed state are

$$\mathbf{r}_T = \mathbf{r}, \quad \mathbf{v}_T = -\mathbf{v}. \qquad \ldots\ldots(5.34)$$

If the momentum is linear in $\mathbf{v}$, then it too will reverse in sign:

$$\mathbf{p}_T = -\mathbf{p}. \qquad \ldots\ldots(5.35)$$

All Poisson bracket relationships are reversed in sign if we make the substitution $(\mathbf{r}, \mathbf{p}) \rightarrow (\mathbf{r}_T, \mathbf{p}_T)$. The transformation is therefore not canonical, but it is particularly simple nevertheless: it is sometimes called an anticanonical transformation. If H_T is obtained from H by these substitutions, then in the new variables the canonical equations of motion become

$$\dot{x}_T = -\frac{\partial H_T}{\partial p_{xT}}, \quad \dot{p}_{xT} = \frac{\partial H_T}{\partial x_T}, \qquad \ldots\ldots(5.36)$$

i.e. again with the signs reversed. The signs may be brought back if $\mathbf{r}$ is considered as a function of $-t$ instead of t. Hence if H_T is the same function of $\mathbf{r}_T$ and $\mathbf{p}_T$ as H is of $\mathbf{r}$ and $\mathbf{p}$, then if $\mathbf{r}(t)$ is a solution of the equations of motion, so also is $\mathbf{r}_T(-t)$. This expresses the concept of reversibility in classical particle mechanics.

In single particle quantum mechanics, corresponding to a state $|\Psi\rangle$, we would like to define a time reversed state $|\Psi_T\rangle$ so that

$$\langle\Psi_T|\mathbf{r}|\Psi_T\rangle = \langle\Psi|\mathbf{r}|\Psi\rangle,$$
$$\langle\Psi_T|\mathbf{v}|\Psi_T\rangle = -\langle\Psi|\mathbf{v}|\Psi\rangle. \qquad \ldots\ldots(5.37)$$

The mapping of states on to the corresponding reversed states cannot be obtained by a unitary transformation; this can be seen in particular if the momentum is just a multiple of $\mathbf{v}$, by taking the expectation value of the commutation relation

$$p_x x - x p_x = -i \qquad \ldots\ldots(5.38)$$

for a state of definite momentum. In order to satisfy (5.37), the corresponding reversed state must have the opposite momentum, and then it is seen that the sign of the expectation value of the commutator must be reversed by time reversal

† This possibility was actually anticipated by Wick, Wightman and Wigner (1952). Its importance for the interpretation of recent experiments was first emphasized by Landau (1957).

of the state. Indeed, by the correspondence principle, the sign of all commutators must be reversed, because of their connection with Poisson brackets. This cannot be achieved by a unitary transformation.

For obvious reasons, however, we wish that

$$|\langle\Psi_T(2)|\Psi_T(1)\rangle|^2 = |\langle\Psi(2)|\Psi(1)\rangle|^2, \qquad \ldots\ldots(5.39)$$

for the right-hand side is the probability of finding the system in state 2 if it is known to be in state 1, and we would not wish such probabilities to be altered when both states are time reversed. Since the possibility

$$\langle\Psi_T(2)|\Psi_T(1)\rangle = \langle\Psi(2)|\Psi(1)\rangle$$

would if generally true imply that the mapping was unitary, which we have seen is unacceptable, we adopt instead the postulate

$$\langle\Psi_T(2)|\Psi_T(1)\rangle = \langle\Psi(2)|\Psi(1)\rangle^* = \langle\Psi(1)|\Psi(2)\rangle. \qquad \ldots\ldots(5.40)$$

A transformation with this property is called anti-unitary.

There are two ways of doing this. In the first place, we could map a complete set of states $|n\rangle$ on to any other complete set $|n'\rangle$, and map a general state

$$|\Psi\rangle = \Sigma c_n|n\rangle,$$

on to

$$|\Psi'\rangle = \Sigma c_n^*|n'\rangle. \qquad \ldots\ldots(5.41)$$

Alternatively, we could map the set of kets $|n\rangle$ on to the set of bras $\langle n'|$, and $|\Psi\rangle$ on to the bra

$$\langle\Psi'| = \Sigma c_n\langle n'|. \qquad \ldots\ldots(5.42)$$

The end-results of (5.41) and (5.42) are related by Hermitian conjugation. We shall normally adopt (5.41), but where appropriate we shall indicate corresponding results for the other alternative. The relation (5.40) will remain unaltered if the transformation (5.41) or (5.42) is followed by a unitary transformation. The most general anticanonical transformation will therefore have the form

$$|\Psi'\rangle = R|\Psi\rangle^*, \qquad \ldots\ldots(5.43)$$

where R is a general unitary transformation and

$$|\Psi\rangle^* = \Sigma c_n^*|n\rangle. \qquad \ldots\ldots(5.44)$$

It is important to realize that the definition (5.44) *depends on a definite choice of basis vectors* $|n\rangle$. A similar remark holds if (5.42) is chosen as the basic transformation.

If the physical meaning of the anti-unitary transformation is to remain unaltered under a change of basis described by a unitary transformation U, then it can be seen that R must become URU^T in the new basis,† instead of the more familiar URU^{-1}. This is similar to the transformation of the $\mathscr{C}$ matrix on change of representative of the Dirac γ matrices, described in § 5.2.

The definition of the time reversal is completed by choosing R in (5.43), in any convenient representation, so that (5.37) is valid. A suitable choice of R is that which gives

$$R\mathbf{r}^*R^{-1} = \mathbf{r}, \quad R\mathbf{p}^*R^{-1} = -\mathbf{p} \qquad \ldots\ldots(5.45)$$

† For $U|\Psi'\rangle = UR|\Psi\rangle^* = UR(U^\dagger U|\Psi\rangle)^* = URU^T(U|\Psi\rangle)^*$.

where $\mathbf{r}^*$ is the operator whose matrix elements *in the chosen representation* are the complex conjugates of those of $\mathbf{r}$.† This defines R up to an arbitrary phase factor.

The time reversed operator $\mathcal{O}_T$ corresponding to any general operator $\mathcal{O}$ is now defined to be‡

$$\mathcal{O}_T = R\mathcal{O}^* R^{-1}, \qquad \ldots\ldots(5.46)$$

so that in particular

$$\mathbf{r}_T = \mathbf{r}, \quad \mathbf{p}_T = -\mathbf{p}. \qquad \ldots\ldots(5.47)$$

The complete definition (5.46) with R defined by (5.45) no longer depends on the choice of representation.

It is sometimes convenient to define a time reversal operator T such that

$$|\Psi_T\rangle = T|\Psi\rangle, \qquad \ldots\ldots(5.48)$$

but great care is needed in doing so, since T is not a linear operator (because of the complex conjugation). It follows that

$$\langle\Psi(2)|\,T^\dagger T\,|\Psi(1)\rangle = \langle\Psi(2)|\Psi(1)\rangle^*, \qquad \ldots\ldots(5.49)$$

and

$$T\mathcal{O}\,T^{-1} = \mathcal{O}_T. \qquad \ldots\ldots(5.50)$$

If we now consider time dependence of states in the Schrödinger picture, from

$$i\frac{\partial}{\partial t}|\Psi(t)\rangle = H|\Psi(t)\rangle, \qquad \ldots\ldots(5.51)$$

we deduce

$$-i\frac{\partial}{\partial t}|\Psi_T(t)\rangle = RH^*R^{-1}|\Psi_T(t)\rangle. \qquad \ldots\ldots(5.52)$$

If

$$RH^*R^{-1} = H \qquad \ldots\ldots(5.53)$$

then the Hamiltonian is invariant under time reversal, and if $|\Psi(t)\rangle$ is a dynamically possible state of the system then so is $|\Psi_T(-t)\rangle$. In the Heisenberg picture, from (5.53) it follows that if $\mathbf{r}(t), \mathbf{p}(t)$ satisfy the canonical equations of motion, so also do $\mathbf{r}_T(-t), \mathbf{p}_T(-t)$.

If one takes instead of (5.41) the point of view of (5.42), then, for example,

$$(\mathbf{r}|\Psi\rangle)_T = \langle\Psi|\mathbf{r}_T{}'\rangle = \langle\Psi|\mathbf{r}_T.$$

In this case, it is easily seen that the order of all operators must be reversed:

$$(ab|\Psi\rangle)_T = \langle\Psi|b_T{}'a_T{}'.$$

† This comes from

$$\begin{aligned}\langle\Psi_T|\mathbf{r}|\Psi_T\rangle &= \sum_{m,n} c_m c_n^* \langle m|R^{-1}\mathbf{r}R|n\rangle,\\ &= \sum c_m c_n^* \langle n|R^T\mathbf{r}^*(R^T)^{-1}|m\rangle,\\ &= \sum c_m c_n^* \langle n|R\mathbf{r}^*R^{-1}|m\rangle,\\ &= \sum c_m c_n^* \langle n|\mathbf{r}|m\rangle = \langle\Psi|\mathbf{r}|\Psi\rangle.\end{aligned}$$

The second step comes from transposition of $R\mathbf{r}^*R^{-1} = \mathbf{r}$, together with the Hermitian character of $\mathbf{r}$.

‡ This definition implies that $(\mathcal{O}|\Psi\rangle)_T = \mathcal{O}_T|\Psi_T\rangle$, and that

$$\langle\Psi_T(2)|\mathcal{O}_T|\Psi_T(1)\rangle = \langle\Psi(2)|\mathcal{O}|\Psi(1)\rangle^*,$$

quite generally for any operator $\mathcal{O}$, whether Hermitian or not. As a simple example, if $\mathcal{O}$ is a complex c-number c, $\mathcal{O}_T = c^*$, and

$$\begin{aligned}\langle\Psi_T(2)|\mathcal{O}_T|\Psi_T(1)\rangle &= c^*\langle\Psi_T(2)|\Psi_T(1)\rangle = c^*\langle\Psi(2)|\Psi(1)\rangle^*\\ &= (c\langle\Psi(2)|\Psi(1)\rangle)^* = \langle\Psi(2)|\mathcal{O}|\Psi(1)\rangle^*.\end{aligned}$$

If one takes this point of view, it is natural to redefine the time reversed operators as the Hermitian conjugates of those defined as in (5.46): this has the merit that whereas a c-number is replaced by its complex conjugate if time reversed according to

$$c_T = Rc^* R^{-1} = c^*,$$

it is left unaltered if this is followed by Hermitian conjugation.

In the single particle case (for a particle without spin) it is convenient to notice that in a representation with $\mathbf{r}$ diagonal and such that $\mathbf{p}$ is proportional to $\mathbf{v}$, T may be unity or a constant phase factor.†

We can now proceed to discuss time reversal for quantized fields by means of the now familiar correspondence arguments. A state in which there is a particle at $\mathbf{r}$ can be written‡

$$\phi^\dagger(\mathbf{r})|0\rangle, \qquad \ldots\ldots(5.54)$$

and we have just seen that the corresponding time reversed single particle state is (apart from an arbitrary phase) identical with this. That is,

$$(\phi^\dagger(\mathbf{r})|0\rangle)_T = e^{-i\vartheta}\phi^\dagger(\mathbf{r})|0\rangle. \qquad \ldots\ldots(5.55)$$

If we postulate that $|0_T\rangle = |0\rangle$, and suitably generalize (5.54) to many particle states, we find on using the footnote to (5.46) that

$$\phi_T(\mathbf{r}) = e^{i\vartheta}\phi(\mathbf{r}) \qquad \ldots\ldots(5.56)$$

and

$$\dot{\phi}_T(\mathbf{r}) = -e^{i\vartheta}\dot{\phi}(\mathbf{r}). \qquad \ldots\ldots(5.57)$$

All c-numbers are to be replaced by their complex conjugates. (If we had adopted the alternative rule of mapping kets on to bras, then we would have had $\phi_T = \phi^\dagger$, $\dot{\phi}_T = -\dot{\phi}^\dagger$, the orders of all factors reversed, and c-numbers left alone.) Equations (5.56) and (5.57) are valid in the Schrödinger picture, or at any given time in the Heisenberg picture. If $\phi(\mathbf{r}, t)$ satisfies the Heisenberg equations of motion with Hamiltonian H, then $\phi_T(\mathbf{r}, -t)$ satisfies the Heisenberg equations of motion with Hamiltonian $H_T = RH^* R^{-1}$.

If time reversal is applied to the charge and current operators, it is found that the charge is unaltered but the current is reversed in sign. If the electromagnetic interactions are to remain invariant under time reversal, it follows that the vector potential must change sign and the scalar potential must remain unaltered. Equivalently, $\mathbf{E}$ must be unaltered, and $\mathbf{H}$ must be reversed.

† This depends on a choice of relative phases of the basic states $|\mathbf{r}\rangle$. If these phases are changed by the unitary transformation which maps $|\mathbf{r}\rangle$ on to $\exp(if(\mathbf{r}))|\mathbf{r}\rangle$, then R can no longer be a constant phase factor, but must be proportional to $\exp(-2if(\mathbf{r}))$. The physical reason for this is that the canonical transformation has changed $\mathbf{p}$ into $\mathbf{p}' = \mathbf{p} - \nabla f$, which is no longer directly proportional to the velocity, and which therefore must not merely be reversed in sign on time reversal. This is discussed by Morpurgo, Touschek and Radicati (1954). Similarly if one transforms to an angular momentum representation, the correct time reversal transformation will depend on the phases chosen for the spherical harmonics (see, for example, Biedenharn and Rose 1953, or Edmonds 1957).

‡ This is a very considerably oversimplified interpretation of the state (5.54), for a relativistic theory, as is discussed for example by Wightman and Schweber (1955). We could avoid the difficulties by using states containing particles of definite momentum, as is done, for example by Morpurgo *et al.* (1954). The results obtained are identical with those we shall obtain, and we feel our approach, while less rigorous, is perhaps more easily followed.

For a particle with spin, it is necessary to choose the single particle R so that the spin is reversed, since spin has the character of a velocity. For a vector field, this leads to the field theory transformation

$$(\phi_0)_T = -e^{i\delta}\phi_0,$$
$$(\phi_k)_T = e^{i\delta}\phi_k, \qquad \ldots\ldots(5.58)$$

and quite generally, for a component $\phi_{\lambda\mu\nu\ldots}$ of a tensor field

$$(\phi_{\lambda\mu\nu\ldots})_T = (-)^{n_1} e^{i\delta}\phi_{\lambda\mu\nu\ldots}, \qquad \ldots\ldots(5.59)$$

where n_1 is the number of indices 0 among $\lambda, \mu, \nu, \ldots$. For a spinor field, the transformation is

$$\psi_T = e^{i\delta} i\gamma_0\gamma_5\psi \qquad \ldots\ldots(5.60)$$

where δ is an arbitrary phase angle.

It is easily verified that time reversal as defined by (5.46), together with equations (5.54)–(5.55), leaves the total charge of the state unaltered, and reverses the direction of the current. This is called *Wigner time reversal*, and was first discussed by Wigner (1932). It is also possible to define a time reversal which leaves the current invariant and reverses the charge, and indeed is Wigner time reversal either preceded by or followed by charge conjugation. This is *Schwinger time reversal* (Schwinger 1951). One should distinguish between Schwinger time reversal and the technique of mapping kets on to bras, also due to Schwinger (1951), which can be used for either transformation.

For a Hermitian field, in order to preserve the Hermitian property it is necessary to choose $e^{i\delta} = \pm 1$. We have already seen, for example, that for Wigner time reversal, the electromagnetic four vector potential A_μ must transform according to (5.58) with $e^{i\delta} = -1$, if the electromagnetic interactions are to be invariant.

Because of the complex conjugation of c-numbers associated with time reversal, $TP\phi P^{-1}T^{-1} = (\phi_P)_T$ may not be the same as $PT\phi T^{-1}P^{-1} = (\phi_T)_P$, so that P and T need not commute: if, however, the phase in the parity transformation is suitably chosen, then P and T will commute. Similarly, whether C and T commute or not depends on the choice of phase associated with charge conjugation.

Since T is not a linear operator, it is not possible to regard it as an observable in the usual sense of quantum mechanics. There is therefore little value in the concept of 'time parity', as an analogue of space parity. It may, perhaps, be useful, however, to use 'intrinsic time parity' to describe the phase factors in definitions such as (5.56).

Applications of invariance under time reversal to problems of detailed balancing and the probabilities of inverse processes are too well known to need discussion here (see, for example, Blatt and Weisskopf 1952). Invariance may also imply relations between the phases of complex coupling constants, as in beta decay (cf. Biedenharn and Rose 1951, Tolhoek and de Groot 1951), and may perhaps rule out the simultaneous occurrence of two different interactions (cf. Coester 1951).

Another application of time reversal, due to Wick, Wightman and Wigner (1952), is the derivation of a *superselection rule*. A superselection rule is said to hold between two sets of states if no observable quantity has a non-zero matrix element between any state of the one set and any state of the other. The argument is as follows. Although the operator T is not linear, T^2 is linear and leaves inner products of state vectors unaltered. T^2 is therefore a unitary operator.

From (5.59) and (5.60), it is clear that

$$T^2\phi_{\lambda\mu\nu\ldots}T^{-2} = \phi_{\lambda\mu\nu\ldots} \qquad \ldots\ldots(5.61)$$

for a tensor operator, and

$$T^2\psi T^{-2} = -\psi \qquad \ldots\ldots(5.62)$$

for a spinor operator. It follows that if $|J\rangle$ is any state with total angular momentum $J(J+1)$, then

$$T^2|J\rangle = (-)^{2J}|J\rangle. \qquad \ldots\ldots(5.63)$$

If we believe that double application of time reversal should leave the physical meaning of the theory invariant, it follows that there can exist no observable quantities which have non-zero matrix elements connecting states of integral and half odd integral spin. There is therefore a superselection rule between these two sets of states. We have already implicitly assumed the existence of superselection rules in addition to this one in our discussion of intrinsic parity—we shall return to this subject in § 5.5.

Recently, time reversal has gained in importance on account of the *CPT* theorem, to be discussed in § 5.4.

Among references for further reading, we would mention particularly Wigner (1932), Schwinger (1951), Watanabe (1951), Morpurgo, Touschek and Radicati (1954), Bell (1955) and Pauli (1955).

5.4. *The CPT Theorem*

There always exists some anti-unitary transformation which leaves a (Hermitian) Hamiltonian invariant, for if H is Hermitian, H^* and H have the same eigenvalues, so that there must exist a unitary operator U where

$$UH^*U^{-1} = H \qquad \ldots\ldots(5.64)$$

(cf. Morpurgo, Touschek and Radicati 1954).

It was shown by Lüders (1954) that for a theory which is invariant under proper Lorentz transformations (i.e. without reflections), with only local interactions, and with the usual commutation rules for field quantities, a particular unitary operator with the property (5.64) is

$$U = CPR, \qquad \ldots\ldots(5.65)$$

where R is the unitary operator associated with time reversal, as defined in § 5.3, and of course it will generally be necessary to make a special choice of the phase factors in the definitions of ϕ_P and ϕ_C.

The neatest proof is that given by Lüders (1957), which we shall sketch very briefly. We first note that in a four-dimensional Euclidean space it is possible to find a *rotation* which reverses the directions of all four axes simultaneously, so that if a quantity is invariant under general rotations it is in particular invariant under this reversal of axes. Under this transformation, a tensor quantity $\phi_{\lambda\mu\ldots}(x)$ with n indices will be transformed into $(-)^n\phi_{\lambda\mu\ldots}(-x)$, and a spinor quantity $\psi(x)$ will be transformed into $\gamma_5\psi(-x)$.

In view of the close connection between tensors and spinors in space–time and in a four-dimensional Euclidean space, discussed in § 3.3, it would not be

surprising if a quantity which is invariant under rotations in space and Lorentz transformations were also invariant under the formal inversion defined by

$$\phi_{\lambda\mu\ldots}(\mathbf{r}, t) \rightarrow (-)^n \phi_{\lambda\mu\ldots}(-\mathbf{r}, -t),$$
$$\psi(\mathbf{r}, t) \rightarrow \gamma_5 \psi(-\mathbf{r}, -t). \qquad \ldots\ldots(5.66)$$

It is found that invariant quantities built up from tensor fields are in fact invariant under (5.66), but that quantities built from $2n$ spinors are multiplied by $(-1)^n$. The different behaviour of quantities built from spinors in space–time and corresponding quantities in a Euclidean space is the only consequence of the differences in the metric between the two types of space.

We wish to apply this result to the Hamiltonian, which is the integral over all space of a Hamiltonian density $H(x)$. (The fact that $H(x)$ is a component of a second-rank tensor rather than a scalar causes no trouble.) Under the transformation (5.66) a term in this density with $2n$ spinor factors is transformed into $(-)^n H(-x)$. The reversal in sign of the space coordinates is unimportant for discussing the invariance of the Hamiltonian, since they are integration variables running over all space. The same is not true for the time variable, however.† A similar difficulty was overcome in the case of time reversal by reversing the signs of time derivatives of the fields, but not changing the sign of t in the arguments.‡ We adopt the same solution in the present case. The result is that a term constructed from $2n$ spinors in the total Hamiltonian, say

$$\mathscr{H}(t) = \int H(x)\, d^3\mathbf{r},$$

is transformed into $(-1)^n \mathscr{H}(t)$.

Now it is well known that to make a consistent theory it is necessary that scalar, or in general tensor, fields must commute for space-like separations of their field points, whereas spinor fields must anticommute. Also, in order to avoid difficulties concerning 'zero point' quantities, it is customary to symmetrize all physical operators with respect to scalar or tensor fields and their Hermitian conjugates and to antisymmetrize for spinor fields. This has been done, for example, in the expressions for charge and current given in equations (5.16), (5.23) and (5.24). It can then be seen that the possible change of sign in quantities built from spinors can be removed by reversing the order of all operators. A very general proof of the results of this and the preceding paragraphs has been given by Pauli (1955) and by Lüders (1957).

We have reached the stage of seeing that the Hamiltonian is invariant under the transformation which we have described, accompanied by reversal of the order of all operators. Because of this reversal of order, the transformation cannot be unitary, but it might be an anti-unitary transformation of the type where kets are mapped on to bras. To change to a transformation where kets are mapped on to kets, we need only perform a Hermitian conjugation, which will also

† We are here deriving the invariance principle independently of the dynamics of the system. The fact that the apparent time dependence of the Hamiltonian proves to be spurious because of the dynamics is irrelevant for our argument.

‡ It was precisely this that forced time reversal to be an anti-unitary rather than a unitary transformation.

leave the quantity invariant provided it is a Hermitian operator. The overall transformation is then

$$\left.\begin{aligned}\phi_{\lambda\mu\ldots}(\mathbf{r},t) &\rightarrow (-)^n \phi_{\lambda\mu\ldots}^{\dagger}(-\mathbf{r},t),\\ \psi(\mathbf{r},t) &\rightarrow -\gamma_5 \psi^{\dagger}(-\mathbf{r},t),\end{aligned}\right\} \qquad \ldots\ldots(5.67)$$

all c-numbers replaced by their complex conjugate.

The last, and also the absence now of any reversal of order of operators, is a consequence of the Hermitian conjugation. The minus sign in the spinor transformation arises because of the purely imaginary character of γ_5 in a Majorana representation. It is now easy to see, by reference to §§5.1, 5.2 and 5.3, that with suitable choices of the phases in the transformations, the transformation (5.67) is just *CPT*. This completes the proof of the theorem.

The importance of the *CPT* theorem lies partly in the connection it gives between processes involving antiparticles, and time reversed processes involving particles, and the various theorems which can be derived therefrom (see, for example, Lee, Oehme and Yang 1957). The other main importance of the theorem is in testing for invariance under charge conjugation, for it is much easier to devise experiments to test this invariance principle indirectly by investigating time reversal invariance and using the *CPT* theorem than by attempting a direct test.

Of course, it may be that we shall eventually be forced to adopt a theory which does not satisfy the conditions of the *CPT* theorem. Even so, provided a Hermitian Hamiltonian exists, there will be some anti-unitary transformation which leaves it invariant, and if this can be identified then it may prove as useful in the new theory as the *CPT* transformation has done in the present theories.

5.5. *Tests of Invariance and Determination of Intrinsic Parities*

We have seen in §5.1 that if it is possible at all to choose phases for the various fields in the parity transformation so as to leave the Hamiltonian invariant, some information will be obtained concerning intrinsic parities of the particles. In this section, we shall discuss how one may test experimentally for invariance properties, and the extent to which an assignment of relative parities is actually possible.

The most obvious test of space-inversion invariance would be a direct test of the existence of a parity operator P which is a constant of the motion. One consequence of this would be the existence of parity selection rules: if, for example, a particle A decays into particles B and C, then the total parity, both orbital and intrinsic, of the final system of two particles must be the same as the parity of the initial system, which in the centre of mass frame of reference is just the intrinsic parity of particle A. It was the apparent breakdown of such a selection rule in the decay of K mesons into two and three π mesons which led Lee and Yang (1956) to investigate other possible tests for invariance under space inversion.

The apparent breakdown of selection rules is not unambiguous evidence for absence of invariance under the parity operation, as it is always conceivable that the initial systems which give rise to two decay modes with opposite parity are not identical. However, any experiment which detects macroscopically an intrinsic asymmetry between left-handed and right-handed screw senses which

cannot be attributed to the experimental arrangement, is conclusive proof of a breakdown of invariance under space inversion because of our correspondence principle definition of the parity transformation. The classic experiment of Wu *et al.* (1957) is of this type. It is now well known that such asymmetries have been observed in beta decay, muon decay, pion decay, K meson decay (into a muon and a neutral particle) and Λ decay. The formal quantum mechanical argument is briefly sketched below.

A typical experiment consists of the preparation of suitable initial states $|\mathrm{i}\rangle$ defined by various momenta and angular momenta, etc., followed by determination of the probabilities of finding the system later in various final states $|\mathrm{f}\rangle$ defined by momenta and angular momenta of the final particles. We shall denote this probability by $W(\mathbf{p}, \mathbf{s})$ where $\mathbf{p}$ stands generically for all the momenta, both initial and final, and $\mathbf{s}$ for all the angular momenta. Quantum mechanics gives a procedure for calculating W from $|\mathrm{i}\rangle, |\mathrm{f}\rangle$ and the Hamiltonian H. Furthermore, according to general principles of quantum theory, we shall get exactly the same result if we apply the same procedure to $U|\mathrm{i}\rangle, U|\mathrm{f}\rangle$ and UHU^{-1}, where U is any unitary operator. If we take for U the parity operator P, then $P|\mathrm{i}\rangle$ differs from $|\mathrm{i}\rangle$ and $P|\mathrm{f}\rangle$ from $|\mathrm{f}\rangle$ only in that all the momenta, angular momenta, etc., are replaced by the corresponding quantities after space inversion according to the classical prescription ; that is, momenta are reversed and angular momenta unaltered. Hence

$$W(\mathbf{p}, \mathbf{s}) = W_P(-\mathbf{p}, \mathbf{s}), \qquad \ldots\ldots(5.68)$$

where W_P is to be calculated in the same manner as W, but using $H_P = PHP^{-1}$ instead of H. If $H_P = H$, then W_P is the same function of its arguments as W, and therefore

$$W(\mathbf{p}, \mathbf{s}) = W(-\mathbf{p}, \mathbf{s}). \qquad \ldots\ldots(5.69)$$

In order to test (5.69), it is necessary only to examine how $W(\mathbf{p}, \mathbf{s})$ depends on quantities such as $\mathbf{s} \cdot \mathbf{p}$ which change sign on performing a classical space inversion. Any asymmetry in W is conclusive proof of the impossibility of choosing the phases in the parity transformation to leave the Hamiltonian invariant.

We have here ignored the fact that some fields, such as the electromagnetic field, may be observable, so that the phase in their parity transformation is defined by their correspondence principle limit. The argument is easily generalized by suitably restricting the choice of phases and by including such fields, where appropriate, among the arguments of W.

It is worth commenting further on the foundations of the argument. It depends only on (i) the unitarity of the space inversion operator and general principles of quantum mechanics, and (ii) the use of the correspondence principle in defining the parity operator. The first assumption cannot be given up without destroying the correspondence principle connection between unitary transformations and canonical transformations, or without rejecting the normal probability interpretation of the theory. The second is necessary if the transformation is reasonably to be called space inversion. Any transformation which leaves the Hamiltonian invariant in a way which does not forbid asymmetries between left- and right-handed screw senses in the results of experiments must violate either (i) or (ii) above.

A similar argument holds for charge conjugation. Here one must include in the specification of the initial and final states a specification of whether particles or antiparticles of the various types are involved. (It is well known, for example, that the signs of certain interference effects in beta decay depend on whether an electron or a positron is emitted.) Invariance of the Hamiltonian under charge conjugation ensures that the probability of any process is identical with the probability of the corresponding process where all particles have been replaced by their antiparticles. Direct tests of this invariance property are difficult, for various reasons such as the unavailability of antinuclei. The most hopeful chances of testing charge conjugation invariance directly are in the decays of pions, muons and K mesons. If the *CPT* theorem is valid, then any asymmetry associated with breakdown of *C* invariance must also be associated with a breakdown of either *P* or *T* invariance.

Time reversal is rather more complicated, since from the nature of the transformation it links a process leading from a state $|\mathrm{i}\rangle$ to a state $|\mathrm{f}\rangle$ with a reversed process leading from a reversed state $|\mathrm{f_T}\rangle$ to a reversed state $|\mathrm{i_T}\rangle$ (cf. Blatt and Weisskopf 1952). We wish, however, to have a connection between processes in which the roles of initial and final states are not interchanged. This is easily achieved if it is valid to treat the process only in the lowest non-vanishing order of perturbation theory. In this case, it is possible to show that if $H_T = H$ then the probability of the process leading from $|\mathrm{i}\rangle$ to $|\mathrm{f}\rangle$ is the same as the probability of the process leading from $|\mathrm{i_T}\rangle$ to $|\mathrm{f_T}\rangle$. If initial and final states are free particle states, this implies

$$W(\mathbf{p}, \mathbf{s}) = W(-\mathbf{p}, -\mathbf{s}), \qquad \ldots\ldots(5.70)$$

i.e. the probability of any process is the same as for the process in which all velocities are reversed. If the initial and final states are not free particle states but differ from free particle states by a phase factor, $\mathrm{e}^{i\delta}$ say, where we use δ generically to denote all such phase-shifts, then by the definition of time reversal, δ must change sign. Hence instead of (5.70) we have

$$W(\mathbf{p}, \mathbf{s}, \delta) = W(-\mathbf{p}, -\mathbf{s}, -\delta). \qquad \ldots\ldots(5.71)$$

Such phase-shifts arise in the final electron state in beta decay processes, due to the effect of Coulomb forces, and their presence makes an experimental check on reversibility difficult unless very light nuclei—preferably the neutron—are used.

The present experimental position is that the interactions responsible for weak decay processes are not invariant under space inversion or charge conjugation, but are probably invariant under time reversal.

The evidence for violation of charge conjugation invariance is based on the *CPT* theorem, whereby a breakdown of invariance under space-inversion must be accompanied by a breakdown of invariance under either charge conjugation or time reversal or both. If a term is observed in *W* which would change sign under space inversion but not under time reversal, then this particular asymmetry must be associated with simultaneous breakdown of *C* and *P* invariance. If all the particles are in free states, such a term could be of the type $\mathbf{s}.\mathbf{p}$. If the free particle states are distorted by other forces, then there will be two kinds of terms of this type, such as $\mathbf{s}.\mathbf{p}\cos\delta$ and $\mathbf{s}.\mathbf{p}\sin\delta$, where δ is a phase shift of an actual

particle state from a free state. These terms are respectively associated with simultaneous breakdown of C and P and of T and P invariance. In many situations, as in beta decay for example, the distortion of the free particle states can be reliably estimated, and indicates that the observed asymmetries are too large to be associated entirely with simultaneous violation of T and P invariance.

We shall now discuss what meaning can be given to the concept of intrinsic parity. In the first place, parity cannot be an exact constant of the motion, so that stationary states cannot be states of sharp parity. However, the properties of stationary states are dominated by the strong and electromagnetic interactions which do appear to be invariant separately under the three inversion operations. We may therefore define intrinsic parities by means of the choice of phases necessary to achieve this invariance. Henceforth, we ignore weak interactions in the discussion of the intrinsic parities of particles.

Secondly, we note that whenever a field has a classical correspondence limit, we must choose the phase in the parity transformation to agree with that required to leave the classical theory invariant under space inversion. In practice, the electromagnetic field is the only one to have a correspondence limit which can be realized experimentally, although there appears to be no reason in principle why any field quantized according to Bose statistics should not have a correspondence limit.†

Thirdly, the possibility of defining the relative parity of two states implies the possibility of determining the relative phase of the two states, and this in turn depends on the existence of an observable which has a non-vanishing matrix element between them. If all possible states can be divided into two or more sets such that no observable has non-vanishing matrix elements between any state drawn from one set and any state from any other set, then a superselection rule exists between the sets and it is impossible to determine the relative parities of states belonging to different sets. We have already seen in §5.3 that a superselection rule holds between states of integral and of half odd integral angular momentum. It must also be remembered that even if some observables have non-vanishing matrix elements between two sets of states, if these vanish when all the weak interactions are made to vanish, then from the point of view of determining parities the situation is the same as if a superselection rule held.

The question of determining intrinsic parities, from the theoretical aspect at least, is thus equivalent to determining the superselection rules, which in turn requires a decision about what is and what is not observable. Since all measurements are in the last resort macroscopic and classical, it seems reasonable to regard as observable only those quantities which possess a correspondence principle limit. In practice this limits us to a few constants of the motion such as total energy, momentum, angular momentum, charge, etc., to the electromagnetic field, and to operators related essentially to the particle aspects of the theory, such as numbers of particles of different kinds, and the single particle operators necessary

† The existence of a correspondence limit of a field implies the measurability of the phase of the field, and more particularly the relative phase of different Fourier components. The uncertainty principle then shows that, if the phase of a Fourier component of the field is well defined, then the number of quanta of the field with the corresponding momentum is completely uncertain. This poses no difficulty for particles for which there is no Pauli exclusion principle, that is for particles satisfying Bose statistics, but clearly requires more careful investigation for particles obeying the exclusion principle.

to define their states. *None of these observables has non-zero matrix elements connecting states with different total charge.* This suggests that a superselection rule exists between such states, as was suggested by Wick, Wightman and Wigner (1952).

The impossibility of measuring the relative parities of states of different charge is connected with the invariance of the Hamiltonian under phase transformations of the fields (gauge transformations of the first kind), which, as shown in §4, is intimately connected with the conservation law of electric charge. For if there is a superselection rule between states of different charge, then all physical observables, not merely the Hamiltonian, are invariant under the gauge transformation, so that the phase of the field operators and therefore the parities of particle states are completely unobservable. Now there are other conservation laws which, like charge, are associated with invariance under gauge transformations of the first kind; for example, conservation of baryons, conservation of leptons, and, if weak interactions are ignored, conservation of strangeness, and separate conservation of electrons, muons and neutrinos. If the only observables are, as already suggested, those with experimentally realizable correspondence principle limits, then with each such conservation law there will be associated a superselection rule. This point of view is certainly consistent with all presently available experiments.

We reach the following conclusions regarding the determination of intrinsic parities. The first problem is to establish all conservation laws of the type of charge conservation (i.e. such that each particle carries a fixed unit, depending on the nature of the particle, of the conserved quantity). Then the relative parities of two states can be determined if and only if they have the same value for all the types of conserved 'charge'. If the vacuum has even parity, then the intrinsic parities of all states of zero charge (of all types) can be determined, in particular the electromagnetic vector potential $\mathbf{A}$ is a polar rather than an axial vector. The intrinsic parity of the π^0 can be determined in principle, though this requires measurements of, for example, the polarizations of the emitted γ-rays in the π^0 decay. The relative parities of electron, muon, pion, nucleon, K meson, etc. cannot be determined. In principle it should be possible to determine the relative parities of Σ^0 and Λ^0, but it seems unlikely that any practicable experimental procedure can be invented.

It is not possible to proceed further without additional postulates. If it is assumed that the neutron and proton have even relative parity, then the parity of the charged pions can be determined, and compared with the independent determination of the π^0 parity. It is usually assumed that both charged and neutral pions have the same intrinsic parity, namely -1. Pais and Treiman (1958) have shown how it is then possible in principle to determine the relative parity of Σ^- and Λ^0. It appears to be true for nucleons and pions that intrinsic parities are the same for all particles belonging to the same isobaric multiplet. It is hard to devise methods of deciding whether the rule also holds good for strange particles. Pais (1958b) has shown that it may in fact be a less natural rule for K mesons than for other particles, and has investigated some of the consequences of charged and neutral K mesons having opposite parity.

Relative parities of particles of different strangeness still remain indeterminate unless further postulates are made.

§6. Isobaric Spin (Nucleons and Pions)

6.1. *Nucleons*

The notion of isobaric spin was first formulated in describing how atomic nuclei are built up from neutrons and protons. The masses of the proton and the neutron are quite close to being equal and the electric charge on the proton plays only a secondary role in nuclear structure: it is therefore very natural to think of the units of which nuclei are composed simply as *nucleons* which may exist either in the charged (proton) state or in the uncharged (neutron) state (Heisenberg 1932). This way of thinking becomes even more useful as soon as the idea is taken up that the forces between nucleons are, in part at least, *exchange* forces, associated with the transfer of electric charge from one nucleon to another.

It so happens (Cassen and Condon 1936) that the mathematical formalism best adapted to describe this situation can be taken over without change from the theory of the electron (or nucleon) spin, which is of course a simple special case of the theory of the rotation group, described in §3.1. At first sight there is nothing more than a very formal analogy between the two situations and in the present case there is no fundamental principle involved comparable to the rotational invariance in the case of spin. The analogy consists simply in the following. If, in the case of the spin, we choose the direction of the spatial x_3-axis to be preferred, the state of a particle in which the spin points in the $+x_3$ direction is characterized by $\sigma_3\psi_+ = +\psi_+$ and the state with spin downwards by $\sigma_3\psi_- = -\psi_-$. Thus the spin operator σ_3 can be represented by the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and the state functions for the states of positive spin will have the column matrix

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

as a factor, while those with negative spin will involve

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

For the two states of the nucleon we can similarly define a matrix

$$\tau_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \ldots\ldots(6.1)$$

(using the suffix 3 merely to stress the analogy) and associate the column vectors

$$|\mathrm{p}\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad |\mathrm{n}\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with the proton and neutrons states of the nucleon respectively. We have then $\tau_3|\mathrm{p}\rangle = +|\mathrm{p}\rangle$, $\tau_3|\mathrm{n}\rangle = -|\mathrm{n}\rangle$, and may note that the operator $\frac{1}{2}(1+\tau_3)$ has the eigenvalue 1 for the proton and 0 for the neutron, so that this is proportional to the nucleon charge. In connection with exchange forces we shall further be interested in operators which convert a neutron into a proton and vice versa. The

former process can be thought of as effected by the operator

$$\tau_+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

for we can readily verify that

$$\tau_+|\mathrm{p}\rangle = 0, \quad \tau_+|\mathrm{n}\rangle = |\mathrm{p}\rangle. \qquad \ldots\ldots(6.2)$$

Conversely, if

$$\tau_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

we find

$$\tau_-|\mathrm{p}\rangle = |\mathrm{n}\rangle, \quad \tau_-|\mathrm{n}\rangle = 0, \qquad \ldots\ldots(6.3)$$

so that τ_- is the operator converting a proton into a neutron. If, following the analogy with the spin, we introduce

$$\tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \qquad \ldots\ldots(6.4)$$

we have evidently $\tau_\pm = \frac{1}{2}(\tau_1 \pm i\tau_2)$.

Thus the algebra of spin operators (including the important relations

$$\tau_i\tau_j = i\tau_k \quad (i,j,k \text{ cyclic}))$$

can be used, quite formally, in describing interactions involving nucleons.

This was, in effect, initiated by Heisenberg in his historic first paper on exchange forces, to which we referred above. However, the gain from such a description would be slight, but for the fact that experiment has indicated that the forces between nuclei show certain important symmetries, to the description of which the formalism is remarkably well adapted. We cannot here go into the experimental evidence for these symmetries, referring the reader to existing literature on the subject (cf. Burcham 1955, Segrè 1953). The charge independence hypothesis was first introduced by Breit, Condon and Present (1936), and the isobaric spin formalism proposed by Cassen and Condon (1936)). We simply state that nuclear forces are now known to be not only *charge symmetric*, but over and above this, to a remarkably good accuracy *charge independent*.

The former statement means that to extremely high accuracy the forces acting between two protons (apart from their electromagnetic interactions) are the same as those between two neutrons. The latter statement expresses the more subtle result, that *provided states of similar space and spin symmetry are compared*, the neutron–proton force is also the same as the proton–proton (or neutron–neutron) force. The proviso in the last sentence is very important because the system of one neutron and one proton can have many states which for two like particles are forbidden by Pauli's principle—for example the deuteron ground state, which is symmetric in both space and spin coordinates. The postulate of charge independence makes no statement about the nucleon force in a state such as this, but it does say that any excited state of the deuteron which is antisymmetric on exchange of space and spin coordinates together, should have its exact counterpart for a two-proton or two-neutron system.

Let us demonstrate how the isotopic spin formalism describes these symmetries in the interaction of two nucleons. Before utilizing either the charge symmetry or the charge independence postulate we note that the general character

of a two-body interaction is restricted by the requirements that the two particles appear symmetrically in it and that it conserves electric charge. Since we are considering exchange forces this charge will in general not be fixed on a single nucleon, but it will rather pertain to the system as a whole. Its constancy can be ensured by making the interaction Hamiltonian depend only on combinations of the τ-operators which commute with the total charge, i.e. with $\frac{1}{2}(\tau_3^{(1)}+\tau_3^{(2)}+2)$, where the superfixes refer to the two interacting particles. It is easy to see that the most general two-body interaction energy that is symmetric in the two particles and satisfies this requirement will have the form

$$W = A+\tfrac{1}{2}B(\tau_3^{(1)}+\tau_3^{(2)})+C\tau_3^{(1)}\tau_3^{(2)}+D(\tau_+^{(1)}\tau_-^{(2)}+\tau_-^{(1)}\tau_+^{(2)}) \qquad \ldots\ldots(6.5)$$

in which the four quantities A, B, C and D may still be any functions of the space and spin variables (and possibly the momenta) of the two nucleons. Let us now make W act on the wave function of a proton–proton, neutron–neutron and proton–neutron system successively. We find, in obvious notation,

$$W|\mathrm{p}^{(1)}\,\mathrm{p}^{(2)}\rangle = (A+B+C)|\mathrm{p}^{(1)}\,\mathrm{p}^{(2)}\rangle, \qquad \ldots\ldots(6.6)$$

$$W|\mathrm{n}^{(1)}\,\mathrm{n}^{(2)}\rangle = (A-B+C)|\mathrm{n}^{(1)}\,\mathrm{n}^{(2)}\rangle, \qquad \ldots\ldots(6.7)$$

$$W|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle = (A-C)|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle + D|\mathrm{n}^{(1)}\,\mathrm{p}^{(2)}\rangle. \qquad \ldots\ldots(6.8)$$

We also write down the last formula with the roles of p and n interchanged:

$$W|\mathrm{n}^{(1)}\,\mathrm{p}^{(2)}\rangle = (A-C)|\mathrm{n}^{(1)}\,\mathrm{p}^{(2)}\rangle + D|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle. \qquad \ldots\ldots(6.9)$$

It is now simple to satisfy the postulate of charge symmetry: simply by requiring that $B = 0$. Then

$$W|\mathrm{p}^{(1)}\,\mathrm{p}^{(2)}\rangle = (A+C)|\mathrm{p}^{(1)}\,\mathrm{p}^{(2)}\rangle, \quad W|\mathrm{n}^{(1)}\,\mathrm{n}^{(2)}\rangle = (A+C)|\mathrm{n}^{(1)}\,\mathrm{n}^{(2)}\rangle.$$

The more stringent requirement of charge independence, on the other hand, requires some further discussion. As a preliminary we add and subtract the last two equations, getting

$$W(|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle + |\mathrm{p}^{(2)}\,\mathrm{n}^{(1)}\rangle) = (A-C+D)(|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle + |\mathrm{p}^{(2)}\,\mathrm{n}^{(1)}\rangle), \qquad \ldots\ldots(6.10)$$

$$W(|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle - |\mathrm{p}^{(2)}\,\mathrm{n}^{(1)}\rangle) = (A-C-D)(|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle - |\mathrm{p}^{(2)}\,\mathrm{n}^{(1)}\rangle). \qquad \ldots\ldots(6.11)$$

Now, we have as yet given no precise meaning to the symbols $(|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}\rangle \pm |\mathrm{p}^{(2)}\,\mathrm{n}^{(1)}\rangle)$ and could in principle attach either of these, or either of the unsymmetric expressions, to any neutron–proton state, but we now remember that the postulate of charge independence enjoins us to compare only those n–p states with p–p or n–n states which are antisymmetric in space and spin coordinates. It appears very natural to associate with all those states which have their counterparts in the p–p and n–n system the symmetrical isobaric functions $2^{-1/2}|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}+\mathrm{p}^{(2)}\,\mathrm{n}^{(1)}\rangle$ (the factor $2^{-1/2}$ is introduced for normalization) and the function $2^{-1/2}|\mathrm{p}^{(1)}\,\mathrm{n}^{(2)}-\mathrm{p}^{(2)}\,\mathrm{n}^{(1)}\rangle$ with the other states. If we do so, Pauli's principle can be stated as follows: If the isobaric spin is included among the variables of the system, the total wave function of any two-nucleon state is antisymmetric. We shall not discuss here the fairly obvious fact that this conventional definition can be extended immediately to an n-nucleon system. It is, however, worth stressing that we are simply introducing a new formal convention, not in any way generalizing the physical content of Pauli's principle.

What are the consequences of this step for our formulation of charge independence? Evidently, simply that the energy expression (6.6) now must be put equal to (6.10), i.e. that we must require $D = 2C$. The total force in any state will then be

$$W = A + C(\tau_3^{(1)}\tau_3^{(2)} + 2\tau_+^{(1)}\tau_-^{(2)} + 2\tau_-^{(1)}\tau_+^{(2)}). \qquad \ldots\ldots(6.12)$$

Reverting to the matrices τ_1 and τ_2 this expression appears as

$$W = A + C(\tau_1^{(1)}\tau_1^{(2)} + \tau_2^{(1)}\tau_2^{(2)} + \tau_3^{(1)}\tau_3^{(2)}),$$

or in a symbolic vector notation,

$$W = A + C\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}. \qquad \ldots\ldots(6.13)$$

The neat form of this last expression is striking, but the gain to be had from this way of stating charge independence goes a little beyond questions of neatness. Equation (6.13) is of just the same form in the $\boldsymbol{\tau}$, as one would write down for the ordinary spin $\boldsymbol{\sigma}$ from rotational invariance considerations (and excluding spin–orbit interactions to which there obviously cannot be a counterpart in the isobaric spin case). Consequently, although rotation in spin space has physical meaning, and the analogous rotation in 'isobaric space' has not, the results of calculations using this '*isobaric invariant*' energy in one's interaction Hamiltonian, being identical at every step with what one would calculate if true space was involved, are bound to give isobaric invariant results in all circumstances. For this reason, one can predict with confidence, e.g. the occurrence of isobaric multiplets in nuclear structure analogous to ordinary spin multiplets, the existence of which is guaranteed by any theory which starts off with the assumption of isobaric invariance of its initial equations. Of course, one has to bear in mind in applying this idea to practical situations that both the electric forces between protons and all effects due to the proton–neutron mass difference will obscure the symmetries of a completely isobaric invariant theory. That nevertheless isobaric invariance shows up quite distinctly in nuclear structure is nowadays undisputed. We must leave further discussion to the specialist literature.

6.2. *Pions and Nucleons*

The origin of nuclear forces has, since Yukawa's work, been recognized as lying (at least predominantly) in the interaction of nucleons with mesons and it was a natural development of Yukawa's original idea to seek to derive the isobaric invariance of nuclear interactions from an underlying isobaric invariance of meson theory. Yukawa's original theory was designed to explain the neutron–proton interaction with the aid of the reactions that we now write as

$$\left.\begin{aligned} \mathrm{p} &\rightleftharpoons \mathrm{n} + \pi^+ \\ \mathrm{n} &\rightleftharpoons \mathrm{p} + \pi^- \end{aligned}\right\} \qquad \text{and} \qquad \ldots\ldots(6.14)$$

where $\pi^\pm$ denotes *pions* of either charge. In conventional field theory these processes are described in terms of an interaction Hamiltonian of the form

$$g(\psi_\mathrm{P}^\dagger\,\Omega\psi_\mathrm{N}\,\phi + \psi_\mathrm{N}^\dagger\,\Omega\psi_\mathrm{P}\,\phi^\dagger). \qquad \ldots\ldots(6.15)$$

Here ψ_P and ψ_N are the wave functions or, in the quantized field theory, wave operators of the proton and neutron, $\psi_\mathrm{P}^\dagger$ and $\psi_\mathrm{N}^\dagger$ the Hermitian conjugates of these

operators, ϕ the corresponding quantity for the positive pion and $\phi^\dagger$ its Hermitian conjugate. For the negative pion the roles of ϕ and $\phi^\dagger$ are reversed. We need not for the present purposes specify the nature of the matrix Ω (it is the Dirac matrix $\gamma_0\gamma_5$ in the currently most acceptable form of the theory). An interaction such as (6.15) is clearly extremely unlikely to lead to forces between two protons or two neutrons comparable with those it predicts for a proton interacting with a neutron and it was evident, very early in the development of the meson theory of nuclear forces, that the only hope of justifying charge independence (should there be a need for this—charge independence was by no means well established in those days) was to assume the existence of a neutral pion capable of the reactions

$$\begin{aligned} \mathrm{p} &\rightleftharpoons \mathrm{p} + \pi^0 \\ \mathrm{n} &\rightleftharpoons \mathrm{n} + \pi^0 \end{aligned} \qquad \ldots\ldots(6.16)$$

which would be described by a Hamiltonian term

$$g' \psi_{\mathrm{P}}^\dagger \Omega \psi_{\mathrm{P}} \phi_0 + g'' \psi_{\mathrm{N}}^\dagger \Omega \psi_{\mathrm{N}} \phi_0. \qquad \ldots\ldots(6.17)$$

For this reason the neutral pion was postulated to exist by theorists long before there was any experimental evidence for a massive neutral boson of any description (Yukawa and Sakata 1937, Yukawa, Sakata and Taketani 1938, Fröhlich, Heitler and Kemmer 1938).

To make an interaction of the type (6.17) satisfy the requirement of charge symmetry the same operator Ω in the two terms of (6.17) had naturally to be postulated. In addition the absolute magnitudes of g' and g'' had to be taken equal. As before, however, the form of interaction had to be narrowed down considerably more if charge independence was to be achieved. It turned out that if only *one* kind of neutral pion was to be assumed the isobaric form of the nucleon–meson interaction was uniquely determined. We shall refer to the original literature (Kemmer 1938 b, Yukawa, Sakata, Kobayasi and Taketani 1938) for the analysis and merely quote the result in a convenient form. The interaction derived (putting $\Omega = 1$ for simplicity) was

$$H = g \sum_{\alpha=1}^{3} \psi^\dagger \tau_\alpha \psi \phi_\alpha \qquad \ldots\ldots(6.18)$$

where the matrices τ_α and the nucleon wave operators ψ and $\psi^\dagger$ have the same meaning as before, while the three quantities ϕ_α are defined as follows: ϕ_3 is the wave operator for the neutral pion previously written as ϕ_0, and ϕ_1 and ϕ_2 are connected with the ϕ and $\phi^\dagger$ of (6.15) by $\phi_1 = 2^{-1/2}(\phi + \phi^\dagger)$ and $\phi_2 = -2^{-1/2} i(\phi - \phi^\dagger)$. The operators ϕ_1 and ϕ_2 are thus Hermitian as is ϕ_3. In order to see that the interaction (6.18) is charge independent, it is sufficient to convince oneself that the conservation law which one obtains by postulating the invariance of (6.18) under 'isobaric rotation' which means, as is evident from the formal structure of (6.18), by treating (ϕ_1, ϕ_2, ϕ_3) as an 'isobaric vector', is a physically sensible one.

Now, just as for a vector in physical space the vector transformation here in question can be stated as follows: Put

$$\Phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix},$$

then an infinitesimal rotation about, say, the λ-axis, transforms Φ into

$$\Phi' = (I + i\epsilon\rho_\lambda)\Phi, \qquad \ldots\ldots(6.19)$$

the three ρ_λ being 3×3 matrices satisfying

$$\rho_\lambda \rho_\mu - \rho_\mu \rho_\lambda = i\rho_\nu \quad (\lambda, \mu, \nu \text{ cyclic}) \qquad \ldots\ldots(6.20)$$

and forming the $s = 1$ irreducible representation of the three-dimensional rotation group. With the choice of the column vector just indicated none of the ρ operators will be diagonal, but one can make a transformation so that, for instance, ρ_3 is diagonal. Its eigenvalues will be $1, 0, -1$, and it is simple to verify that the eigenfunctions of ρ_3 belonging to these three eigenvalues are ϕ, ϕ_0 and $\phi^\dagger$ respectively. It is thus clearly possible to define the vector $\boldsymbol{\rho}$ as representing the isobaric spin of the pion and the formal analogy with physical rotations will ensure that both the total isobaric spin of the system $\mathbf{T}^2 = (\boldsymbol{\rho} + \frac{1}{2}\boldsymbol{\tau})^2$ and its third component $T_3 = \rho_3 + \frac{1}{2}\tau_3$ will commute with H and therefore will be conserved in all processes effected by the interaction (6.18).

The latter conservation law is evidently, as previously for nucleons alone, equivalent to the conservation law of electric charge. The former has no such obvious physical interpretation, but once established it leads to most helpful physical conclusions, some of which we shall discuss below. Since the eigenvalues of ρ_3 are, as we have seen, associated with the positive, neutral and negative pions respectively, the relation between charge and third isobaric spin component is different from that for nucleons, ρ_3 itself being here proportional to the charge.†

6.3. *Invariance of Interaction*

It is perhaps instructive to demonstrate in an explicit calculation the invariance of the interaction (6.18) under isobaric rotation. This is best done by use of a formalism based essentially on an idea of Harish-Chandra (1946). As the formalism will also serve to simplify some later calculations, we now permit ourselves a short digression to establish it. To apply transformations like (6.19) to the interaction (6.18) directly it is necessary to introduce the column matrix notation Φ for the meson field operators in (6.18). This can be done by defining three row matrices $\Gamma_1, \Gamma_2, \Gamma_3$ such that

$$\phi_\alpha = \Gamma_\alpha \Phi. \qquad \ldots\ldots(6.21)$$

We can then write

$$H = g\psi^\dagger \tau_\alpha \psi \Gamma_\alpha \Phi. \qquad \ldots\ldots(6.22)$$

It is easy to verify that in fact

$$\Gamma_1 = (\sqrt{\tfrac{1}{2}}\, 0\, \sqrt{\tfrac{1}{2}}), \quad \Gamma_2 = (i\sqrt{\tfrac{1}{2}}\, 0\, -i\sqrt{\tfrac{1}{2}}), \quad \Gamma_3 = (0\,1\,0). \qquad \ldots\ldots(6.23)$$

If we denote by ${}^T\Gamma_\alpha$ the column matrix Hermitian conjugate to Γ_α, we can obviously also write

$$H = g\Phi^\dagger {}^T\Gamma_\alpha \psi^\dagger \tau_\alpha \psi. \qquad \ldots\ldots(6.24)$$

It is actually convenient to express H in the form (6.22) when dealing with the absorption parts of the pion operators and as (6.24), when concerned with pion emission.

† In the extension of the theory that will be discussed in § 7 the reason for this difference in definition will become clear. It is connected with the fact that by postulating (6.18) we are tacitly assuming a further conservation law, that of nucleon number.

It is now not difficult to verify the relations

$$\Gamma_\alpha \Gamma_\beta = \delta_{\alpha\beta} \qquad \ldots\ldots(6.25)$$

from which it follows that if

$$\rho_1 = i(\Gamma_2\Gamma_3 - \Gamma_3\Gamma_2), \quad \rho_2 = i(\Gamma_3\Gamma_1 - \Gamma_1\Gamma_3), \quad \rho_3 = i(\Gamma_1\Gamma_2 - \Gamma_2\Gamma_1), \qquad \ldots\ldots(6.26)$$

the three ρ_λ are 3×3 matrices satisfying (6.20). They are in fact the pion spin matrices, in the representation in which ρ_3 is diagonal.

We also note that

$$\Gamma_\lambda \rho_\mu = i\Gamma_\nu \quad (\lambda, \mu, \nu \text{ cyclic}). \qquad \ldots\ldots(6.27)$$

If then we transform ψ, $\psi^\dagger$ and Φ according to the appropriate spin transformation, we get, e.g. for 'rotation about the 3-axis' by the infinitesimal 'angle' ϵ,

$$\psi' = \psi + \tfrac{1}{2}i\epsilon\tau_3\psi, \quad (\psi^\dagger)' = \psi^\dagger - \tfrac{1}{2}i\epsilon\psi^\dagger\tau_3, \quad \Phi' = \Phi + i\epsilon\rho_3\Phi, \qquad \ldots\ldots(6.28)$$

and hence

$$H' = H + gi\epsilon[\psi^\dagger \tfrac{1}{2}(\tau_\alpha\tau_3 - \tau_3\tau_\alpha)\psi\Gamma_\alpha\Phi + \psi^\dagger\tau_\alpha\psi\Gamma_\alpha\rho_3\Phi] = H, \qquad \ldots\ldots(6.29)$$

which is the required formal proof of invariance.

An isobarically invariant formalism must of necessity have isobarically invariant consequences; it is not necessary to prove independently that the nuclear forces deduced in any consistent manner from the Hamiltonian (6.18) are charge independent. Nevertheless it is worth demonstrating how simply this result follows, at least in the second (first relevant) order of perturbation, by formal manipulation only. In this approximation the expression for the nuclear force will be governed by a product of two matrix elements of H, one involving pion emission, the other pion absorption. The nuclear potential will contain the expression

$$\tau_\alpha^{(1)}\Gamma_\alpha\Gamma_\beta\tau_\beta^{(2)}. \qquad \ldots\ldots(6.30)$$

The first pair of factors refers to the pion and the nucleon absorbing it, the second to the same pion and to the emitting nucleus. Using (6.25) it follows directly that the resulting nuclear force must have $\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}$ as a factor. Similar reasons show that in higher order of perturbations only powers of $\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}$ can enter. The conservation of total isobaric spin in processes effected by the interaction (6.22) having been established, we shall now deduce a few important consequences.

Consider the processes (6.14). Since the single particles on the left-hand side have isobaric spin $\frac{1}{2}$, the total isobaric spin of the states on the right-hand side produced by the interaction (6.18) must also be $T = \frac{1}{2}$, while in general the system of one nucleon and one pion may belong to either $T = \frac{3}{2}$ or $T = \frac{1}{2}$, as shown by the vector sum rule, applied to $\frac{1}{2}\boldsymbol{\tau}$ and $\boldsymbol{\rho}$. Among the states belonging to $T = \frac{3}{2}$ are the states $|\mathrm{p}\pi^+\rangle$ and $|\mathrm{n}\pi^-\rangle$ which obviously cannot be produced starting from a single nucleon state for reasons of charge conservation, but of the systems $(\mathrm{p}\pi^-)$, $(\mathrm{n}\pi^0)$, we know to start with only that $T_3 = -\frac{1}{2}$, and they may thus occur either in a $T = \frac{3}{2}$ or $T = \frac{1}{2}$ state. Let us now operate with (6.24) on the single neutron state. If we represent this state by the column vector

$$|\mathrm{n}\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

we can represent the operation by

$$H|\mathrm{n}\rangle = g'\mathcal{T}_\alpha\tau_\alpha\begin{pmatrix}0\\1\end{pmatrix} = g\left[(\mathcal{T}_1 - i'\mathcal{T}_2)\begin{pmatrix}1\\0\end{pmatrix} - \mathcal{T}_3\begin{pmatrix}0\\1\end{pmatrix}\right]. \qquad \ldots\ldots(6.31)$$

We have written the interaction in the form appropriate to pion creation in which case the $\mathcal{T}$ symbols may be interpreted directly as isobaric column vectors of the pion. By their definition we have

$$\mathcal{T}_1 - i'\mathcal{T}_2 = \sqrt{2}\begin{pmatrix}0\\0\\1\end{pmatrix}, \quad \mathcal{T}_3 = \begin{pmatrix}0\\1\\0\end{pmatrix} \qquad \ldots\ldots(6.32)$$

so that we get

$$H|\mathrm{n}\rangle = g\left[\sqrt{2}\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix} - \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}0\\1\\0\end{pmatrix}\right]$$

$$= g[\sqrt{2}|\,\mathrm{p}\pi^-\rangle - |\,\mathrm{n}\pi^0\rangle]. \qquad \ldots\ldots(6.33)$$

Normalizing correctly, we see that the state

$$\sqrt{\tfrac{2}{3}}|\rangle\mathrm{p}\pi^-|\rangle - \sqrt{\tfrac{1}{3}}|\rangle\mathrm{n}\pi^0|\rangle \qquad \ldots\ldots(6.34)$$

must be the state $T = \frac{1}{2}, T_3 = -\frac{1}{2}$. The combination of $|\rangle\mathrm{p}\pi^-|\rangle$ and $|\rangle\mathrm{n}\pi^0|\rangle$ orthogonal to this state, which is $\sqrt{\frac{1}{3}}|\,\mathrm{p}\pi^-\rangle + \sqrt{\frac{2}{3}}|\,\mathrm{n}\pi^0\rangle$ must therefore belong to $T = \frac{3}{2}$; it is part of the same $T = \frac{3}{2}$ multiplet as $|\rangle\mathrm{p}\pi^+|\rangle$. A check on the calculation results from applying the pion absorbing part of the interaction to a $T = \frac{3}{2}$ state. We have

$$g\tau_\alpha\Gamma_\alpha\left[\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix} + \sqrt{2}\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}0\\1\\0\end{pmatrix}\right] = g\left[\begin{pmatrix}1\\0\end{pmatrix}:(\sqrt{\tfrac{1}{2}} + \sqrt{\tfrac{1}{2}}) - \sqrt{2}\begin{pmatrix}1\\0\end{pmatrix}\cdot 1\right] = 0.$$

$\ldots\ldots(6.35)$

As befits a $T = \frac{3}{2}$ state, this one evidently does not combine with any single nucleon state.

The reader may be left to work out similarly the $T = \frac{1}{2}$ and $T = \frac{3}{2}$ states with $T_3 = +\frac{1}{2}$ (total charge $+e$), and is referred to the textbooks for more detailed discussion.

6.4. *Application to Scattering*

These, so far formal, definitions of isobaric states assume very concrete physical significance in connection with the study of pion–nucleon scattering. The scattering processes one has to consider in principle are

$$\begin{aligned}
&(a)\ \pi^+ + \mathrm{p} \to \pi^+ + \mathrm{p}\\
&(b)\ \pi^+ + \mathrm{n} \to \pi^+ + \mathrm{n}\\
&(c)\ \pi^+ + \mathrm{n} \to \pi^0 + \mathrm{p}\ \text{(exchange scattering)}\\
&(d)\ \pi^0 + \mathrm{p} \to \pi^0 + \mathrm{p}\\
&(e)\ \pi^0 + \mathrm{n} \to \pi^0 + \mathrm{n}\\
&(f)\ \pi^- + \mathrm{p} \to \pi^- + \mathrm{p}\\
&(g)\ \pi^- + \mathrm{p} \to \pi^0 + \mathrm{n}\ \text{(exchange scattering)}\\
&(h)\ \pi^- + \mathrm{n} \to \pi^- + \mathrm{n}.
\end{aligned} \qquad \ldots\ldots(6.36)$$

We have not written down the exchange processes for incident uncharged pions because the table includes the processes inverse to them, so that in any reasonable quantum theory they would not appear as independent unknowns. Of course, the majority of processes listed are not observable in any feasible experiments: π^0 scattering, even on protons, is not a process on which quantitative information can be hoped for and all scattering on neutrons could in practice only be deduced, indirectly and with much theoretical uncertainty, from scattering on, say, deuterons. In the absence of symmetry principles it would evidently be hopeless to attempt to gain a complete picture of pion scattering. The magnitude of the problem is, however, substantially reduced by the application of the charge symmetry principle, which states that processes (*a*) and (*h*), (*b*) and (*f*), (*c*) and (*g*), and (*d*) and (*e*) should be describable by identical equations. Since processes (*a*) (*f*) and (*g*) are capable of direct and accurate observation, the acceptance of the charge symmetry principle makes the problem more tractable, but even so, process (*d*) remains practically unmeasurable and no theoretical relation between the other three need exist. If, however, one makes the further step of analysing experiments on the basis of the hypothesis of charge ***independence***, a much more interesting situation arises. The hypothesis implies that scattering amplitudes, etc. shall be taken to depend only on the value of the total isobaric spin involved in the process, not on the particular T_3 component involved. In formulae this means that the 'processes'

$$|\mathrm{p}\pi^+\rangle \rightarrow |\mathrm{p}\pi^+\rangle' \qquad \ldots\ldots(6.37)$$

and

$$[\sqrt{\tfrac{1}{3}}|\mathrm{p}\pi^-\rangle + \sqrt{\tfrac{2}{3}}|\mathrm{n}\pi^0\rangle] \rightarrow [\sqrt{\tfrac{1}{3}}|\mathrm{p}\pi^-\rangle + \sqrt{\tfrac{2}{3}}|\mathrm{n}\pi^0\rangle]' \qquad \ldots\ldots(6.38)$$

must have precisely the same mathematical description while the 'process'

$$[\sqrt{\tfrac{2}{3}}|\mathrm{p}\pi^-\rangle - \sqrt{\tfrac{1}{3}}|\mathrm{n}\pi^0\rangle] \rightarrow [\sqrt{\tfrac{2}{3}}|\mathrm{p}\pi^-\rangle - \sqrt{\tfrac{1}{3}}|\mathrm{n}\pi^0\rangle]' \qquad \ldots\ldots(6.39)$$

may be entirely unrelated to the others. Also, no transitions from a state with $T = \frac{3}{2}$ to a state with $T = \frac{1}{2}$, or vice versa, can occur. We can therefore state the amplitudes of the observed scattering processes as follows:

$$\langle \mathrm{p}\pi^+|\mathrm{p}\pi^+\rangle = a(\tfrac{3}{2}), \qquad \ldots\ldots(6.40)$$

$$\begin{aligned}\langle \mathrm{p}\pi^- \ \mathrm{p}\pi^-\rangle &= [\sqrt{\tfrac{1}{3}}(\sqrt{\tfrac{1}{3}}\langle \mathrm{p}\pi^-| + \sqrt{\tfrac{2}{3}}\langle \mathrm{n}\pi^0|) + \sqrt{\tfrac{2}{3}}(\sqrt{\tfrac{2}{3}}\langle \mathrm{p}\pi^-| - \sqrt{\tfrac{1}{3}}\langle \mathrm{n}\pi^0|)]|\mathrm{p}\pi^-\rangle \\ &= \tfrac{1}{3}a(\tfrac{3}{2}) + \tfrac{2}{3}a(\tfrac{1}{2}). \qquad \ldots\ldots(6.41)\end{aligned}$$

Similarly,

$$\langle \mathrm{p}\pi^-|\mathrm{n}\pi^0\rangle = \sqrt{\tfrac{2}{9}}[a(\tfrac{3}{2}) - a(\tfrac{1}{2})], \qquad \ldots\ldots(6.42)$$

and also

$$\langle \mathrm{n}\pi^0|\mathrm{n}\pi^0\rangle = \tfrac{2}{3}a(\tfrac{3}{2}) + \tfrac{1}{3}a(\tfrac{1}{2}). \qquad \ldots\ldots(6.43)$$

We see thus that the amplitudes for all three observable processes (as well as for the π^0 scattering) are expressible in terms of but two constants. Experimental observation can therefore not only be described in terms of a quite small number of independent quantities, but is sufficient for a check on the conservation of the charge independence assumption. The observations actually made (in the energy region of 190 MeV) vindicate the charge independence hypothesis in a striking fashion. The ratio of cross sections for the processes (*a*), (*f*) and (*g*) in this energy region is measured to be very close indeed to the ratio 9:1:2. Since the cross sections are proportional to the squares of the scattering amplitudes a glance at (6.40) to (6.42) shows that this is what is obtained from charge independence

theory if one puts $a(\frac{1}{2}) = 0$, the scattering thus appearing to be predominantly due to the $T = \frac{3}{2}$ state—in which a resonance phenomenon occurs not shared by the $T = \frac{1}{2}$ state interaction.

Today the isobaric spin formalism is a well-established and powerful tool of research in pion physics, and one can have little doubt that in spite of the very formal origin of the isobaric spin concept it expresses a deep and as yet very incompletely understood principle, governing the structure of fundamental particles.

In the following chapter we shall see how the application of the same principle to the 'strange particles' leads both to a consolidation and to a widening of these ideas.

§ 7. Isobaric Spin (Strange Particles)

Since 1945 experiments with cosmic rays and high-energy accelerators have shown that an unsuspected wealth of elementary particles exists. These particles have received the generic name 'strange'; not only because their occurrence was unexpected, but also because their properties at first appeared to be anomalous.

7.1. *Associated Production and the Gell-Mann–Nishijima Scheme*

The strange particles are produced in fair abundance. For example, the cross section for strange particle production by 950 MeV pions is about 1 mbn. In fact, an estimate of the strength of the production interaction indicates that it is of roughly the same order of magnitude as that of the pion–nucleon interaction. This immediately raises difficulties in view of the fact that these particles are comparatively long-lived with lifetimes of the order of 10^{-10} second. If, for example, the Λ^0 is produced by the reaction

$$\pi^- + N^+ \rightarrow \Lambda^0 + \pi^0, \qquad \ldots\ldots(7.1)$$

then on very general grounds we should expect the following chain of strong virtual processes to be possible:

$$\Lambda^0 \rightarrow N^+ + \pi^- + \pi^0 \rightarrow N^+ + \pi^-. \qquad \ldots\ldots(7.2)$$

The second link in the chain is the strong Yukawa absorption of the π^0. (7.2) would produce the decay of the Λ^0 with a lifetime of the order of 10^{-23} sec.

One possible resolution of this difficulty was proposed by Fermi and Feynman (unpublished). This supposes that the matrix elements for the reaction (7.1) vary very sharply with energy. They are to be large at the pion energies corresponding to the threshold for Λ^0 production but small at the energies corresponding to the first virtual process in (7.2). A model incorporating this feature would be one that made the hyperons excited states of the pion–nucleon system with high intrinsic spin. The resulting centrifugal barrier would sufficiently inhibit the reaction (7.2) to bring the Λ^0 lifetime into accord with experiment if the Λ^0 had a spin of 13/2. Near threshold this model would favour the associated production of two strange particles, as the angular momentum balance would then be easier. (Only low partial-wave pions are present at these energies.) The process of this type with the lowest threshold is

$$N^0 + N^0 \rightarrow \Lambda^0 + \Lambda^0. \qquad \ldots\ldots(7.3)$$

However, no evidence has been found for this reaction. The acid test of the theory is the measurement of the spin of the Λ^0. There is every indication that this is 1/2 (CERN 1958). It is therefore necessary to find another resolution of the paradox.

This is provided by the suggestion of Pais (1952) that strange particles are never produced singly but only in associated pairs. Thus an interaction of the form (7.1) is not permitted and the production process must be by some such reaction as

$$\pi^- + N^+ \rightarrow \Lambda^0 + K^0. \qquad \ldots\ldots(7.4)$$

This suggestion has been abundantly confirmed by experiment. It is natural to try to interpret this requirement of associated production in terms of a conservation law. One condition that must be satisfied by such a law is that it should exclude the unobserved reaction (7.3).

An important clue to the explanation of pion–nucleon interactions lay in the fact that the particles involved could be divided into two groups, or isobaric multiplets. Within each group particles have approximately the same mass and certain other common properties. Gell-Mann (1953, see also Gell-Mann and Pais 1955 a) and Nishijima (1954) extended this idea to the strange particles and showed that it could be used to furnish a conservation law that explained all the observed phenomena. The obvious multiplets occurring among the strange particles are that containing the Σ^+ and Σ^- hyperons at around 2330 e.m. and that containing the K^+, K^-, K^0 mesons at around 960 e.m. In fact, the Gell-Mann–Nishijima scheme assumed the existence of two particles, Σ^0 and Ξ^0, not then observed, and also that there were two different neutral K-mesons, K^0 and $\bar{K}^0$. The resulting multiplet schemes are shown in tables 2 and 3.

Table 2. Baryon Multiplets

2580 e.m.	Ξ^-	Ξ^0	
2330 e.m.	Σ^-	Σ^0	Σ^+
2180 e.m.		Λ^0	
1830 e.m.		N^0	N^+

Table 3. Meson Multiplets

960 e.m.	K^-	$\bar{K}^0$	K^+
		K^0	
270 e.m.	π^-	π^0	π^+

Among the hyperons the Σ particles form an isobaric triplet with $T = 1$, the Ξ particles and the nucleons two isobaric doublets with $T = \frac{1}{2}$, and the Λ^0 an isobaric singlet, $T = 0$. Among the mesons K^+, K^0 form an isobaric doublet, $T = \frac{1}{2}$, their charge conjugates another isobaric doublet, and the pions a triplet, $T = 1$. However, the strange particle multiplets are characterized by the fact that they do not satisfy the usual relation between charge, third component of isobaric spin, and baryon number

$$Q = T_3 + \tfrac{1}{2}B. \qquad \ldots\ldots(7.5)$$

They are, in fact, *displaced* multiplets in isobaric space and satisfy

$$Q = T_3 + \tfrac{1}{2}B + \tfrac{1}{2}S, \qquad \ldots\ldots(7.6)$$

where S is a quantity specifying their displacement. S is called the *strangeness* of the multiplet and table 4 summarizes the strangeness assignations in the Gell-Mann–Nishijima scheme :

Table 4

$S=-2$	$S=-1$	$S=0$	$S=+1$
Ξ^- Ξ^0	Λ^0	N^0, N^+	K^0, K^+
	Σ^-, Σ^0, Σ^+	π^-, π^0, π^+	
	K^-, $\bar{K}^0$		

All known interactions conserve Q and B and so, from (7.6) any interaction that conserves T_3 must also conserve S. Interactions are divided into three classes according to their intrinsic strengths and conservation laws are associated with each class :

(i) *Strong interactions.* These are the meson–baryon interactions that are characterized by coupling constants of the order of 1. They conserve T and T_3, and so conserve S.

(ii) *Electromagnetic interactions.* These are characterized by the fine structure constant, 1/137. They conserve T_3 (but not T) and so conserve S.

(iii) *Weak interactions.* These are characterized by coupling constants of the order of 10^{-14}. Examples are β-decay interactions, and the interactions causing the decay of strange particles. They conserve neither T nor T_3. Thus S is not conserved but the change in S is supposed to be restricted by the rule

$$|\Delta S| = 1. \qquad \ldots\ldots(7.7)$$

The Gell-Mann–Nishijima scheme (GN scheme) has proved completely successful in providing a qualitative theory of elementary particle interactions. Some of the most important consequences of the scheme are :†

(*a*) *Associated production.* This follows from the fact that nucleons and pions all have zero strangeness. The conservation of S in the strong interactions then implies that strange particles must be produced in pairs. A typical reaction is

$$\pi^- + N^+ \rightarrow \Lambda^0 + K^0. \qquad \ldots\ldots(7.4)$$

We note that (7.3) is forbidden by the scheme.

(*b*) *K-meson properties.* Although K^+ and K^0 can be produced in association with a Λ^0, etc., K^- and $\bar{K}^0$ cannot be produced in this way because they have the same strangeness as Λ^0. They must be produced by a reaction like

$$\pi^- + N^+ \rightarrow K^+ + K^- + N^0 \qquad \ldots\ldots(7.8)$$

which has a higher threshold than (7.4). This explains the fact that K^+s are more abundant in cosmic rays than K^-s.

† Excellent reviews of present knowledge in elementary particle physics have been given by Dalitz (1957) and Gell-Mann and Rosenfeld (1957). We shall be content here to summarize the results most pertinent to our theme.

The only reaction possible for slow K^+s in nuclear matter is scattering. K^-s, however, may be absorbed to form hyperons (often bound in hyper-fragments), e.g.

$$K^- + N^+ \rightarrow \Lambda^0 + \pi^0. \qquad \ldots\ldots(7.9)$$

The K^0 and $\bar{K}^0$ provide an example of a neutral meson with a distinct charge conjugate. This is required on general grounds to forbid the interaction (7.3). The reaction (7.4) together with the Yukawa pion–nucleon interaction makes possible the virtual processes

$$N^0 \rightarrow \pi^- + N^+ \rightarrow \Lambda^0 + K^0, \qquad \ldots\ldots(7.10)$$

and so, also

$$\bar{K}^0 + N^0 \rightarrow \Lambda^0. \qquad \ldots\ldots(7.11)$$

Then if $K^0 = \bar{K}^0$ we should have

$$N^0 + N^0 \rightarrow \Lambda^0 + K^0 + N^0 \rightarrow \Lambda^0 + \Lambda^0. \qquad \ldots\ldots(7.12)$$

Gell-Mann and Pais (1955 b) have pointed out some interesting consequences of the distinctness of K^0 and $\bar{K}^0$. At the time of their original discussion it was thought that the weak interactions were invariant under C. Similar arguments apply, however, if CP is conserved. If we consider the K-meson in its rest frame it is natural to analyse its decay in terms of K_1^0 and K_2^0, where

$$K_1^0 = \frac{K^0 + \bar{K}^0}{\sqrt{2}}, \quad K_2^0 = \frac{K_0 - \bar{K}_0}{\sqrt{2}i}. \qquad \ldots\ldots(7.13)$$

In contrast to K^0 and $\bar{K}^0$, K_1^0 and K_2^0 are eigenstates of CP, with eigenvalues $+1$ and -1 respectively.† Any final states resulting from the decay that are eigenstates of CP are not, therefore, simultaneously accessible to both K_1^0 and K_2^0. In particular only K_1^0 can decay into two pions ($\pi^+ + \pi^-$, or $2\pi^0$) in an S-state since such a state is an eigenstate with $CP = +1$. In general, we should expect that the lifetimes of K_1^0 and K_2^0 would be different and that therefore there would be a short-lived and a long-lived neutral K-meson, each of mixed strangeness. Both short-lived and long-lived decay modes should be observed when K^0s are produced, since

$$K^0 = \frac{K_1^0 - iK_2^0}{\sqrt{2}}. \qquad \ldots\ldots(7.14)$$

Some of the qualitative features of this scheme are retained even if CP is not conserved (Lee, Oehme and Yang 1957). The existence of these short-lived and long-lived neutral K-mesons has been established experimentally (Landé *et al.* 1956).

Some interesting consequences of the mixed strangeness of the long-lived component have been pointed out by Pais and Piccioni (1955).

(*c*) *The* Σ^0 *hyperon.* The scheme postulated the existence of this particle. However, it permits the radiative decay

$$\Sigma^0 \rightarrow \Lambda^0 + \gamma \qquad \ldots\ldots(7.15)$$

which makes Σ^0 highly unstable with a lifetime of about 10^{-20} second. Subsequently this reaction has been detected (Plano *et al.* 1957).

† We have taken relative phases such that $CP\,|\,K^0\rangle = |\,\bar{K}^0\rangle$, $CP\,|\,\bar{K}^0\rangle = |\,K\rangle$.

(*d*) *The* Ξ^- *hyperon.* The rules for the weak interactions are satisfied by the known strange particle decays. In particular, the rule (7.7) requires that the doubly strange Ξ^- should have a cascade decay such as

$$\Xi^- \rightarrow \Lambda^0 + \pi^- \rightarrow N^+ + \pi^- + \pi^-. \qquad \ldots\ldots(7.16)$$

The existence of Ξ^0 is still unconfirmed. (The threshold for cascade production by a reaction like

$$\pi^- + N^+ \rightarrow \Xi^0 + K^0 + K^0 \qquad \ldots\ldots(7.17)$$

is near the maximum bevatron energy.)

Its existence is required by the theory since the Ξ^- has been observed to be produced in association with two K-mesons (Sorrels, Leighton and Anderson 1955). Ξ^- therefore has strangeness -2 and must be a member of a doublet.

The rule (7.7) implies that the weak decays are also characterized by

$$|\Delta T_3| = \tfrac{1}{2} \qquad \ldots\ldots(7.18)$$

for those decays in which all the participating particles have a defined isobaric spin. Gell-Mann and Pais (1955 a) have suggested that the weak decays may also obey the stronger rule

$$|\Delta T| = \tfrac{1}{2} \qquad \ldots\ldots(7.19)$$

One striking piece of evidence in favour of this rule is that the lifetime of K^+ is 500 times longer than that of K^0. It is easy to see that for a spin zero K^+-meson decaying into two pions the pions must be in a $T = 2$ state since their spatial wave function is symmetrical, and so Bose statistics imply that the isobaric wave function is symmetrical also. This state has positive charge and so must be $T = 2$. Thus this decay requires $\Delta T = 3/2$ and its relative slowness is accounted for by the rule (7.19), which of course cannot be exact as electromagnetic corrections will modify it.

(7.19) may also be used to forecast certain branching ratios. It correctly predicts the branching ratio for Λ^0 decay

$$\frac{\Lambda^0 \rightarrow N^0 + \pi^0}{\Lambda^0 \rightarrow N^+ + \pi^-}, \qquad \ldots\ldots(7.20)$$

but is unable to explain the low value of the branching ratio

$$\frac{K_1^0 \rightarrow 2\pi^0}{K_1^0 \rightarrow \pi^+ + \pi^-}. \qquad \ldots\ldots(7.21)$$

In fact, admixtures of $\Delta T = 3/2$ and $\Delta T = 5/2$ may be required to explain the K-meson decays completely (Gell-Mann 1957 a).

An interesting theoretical interpretation of the rule has been suggested by D'Espagnat, Prentki and Salam (1957). This is discussed later.

The scheme assumes that charge independence applies to strong interactions involving strange particles. This would not of course be so if the recent suggestion by Pais (1958 b) that K^+ and K^0 have opposite parity turns out to be the case. The reason for the observed mass spectrum then becomes a critical problem. This is, of course, a more restrictive assumption than that of the conservation of S and T_3. The existence of a multiplet structure for the elementary particles is

strong *prima facie* evidence for it. However, it is interesting to seek further experimental tests. One such simple test has been suggested by Lee (1955). He considers the reactions

$$\mathrm{K}^- + \mathrm{d} \begin{cases} \rightarrow \Sigma^0 + \pi^- + \mathrm{N}^+ \\ \rightarrow \Sigma^- + \pi^0 + \mathrm{N}^+. \end{cases} \qquad \ldots\ldots(7.22)$$

The initial state has $T = \frac{1}{2}$ since the deuteron has isobaric spin zero. Therefore, if isobaric spin is conserved, the $(\Sigma + \pi)$ system, which has charge -1, must have an isobaric spin that can be combined with the proton to give a $T = \frac{1}{2}$ state. This requirement uniquely determines the state of $(\Sigma + \pi)$ to be $T = 1$ and so the branching ratio should be unity. Consequences of isobaric spin conservation in more complicated cases have been considered by Feldman (1956).

7.2. *Mathematical Formulations of the Gell-Mann–Nishijima Scheme*

The outstanding success of the phenomenological relation (7.6) has encouraged the search for other mathematical formulations of the Gell-Mann–Nishijima scheme that may, perhaps, throw some light on the mysteries of isobaric spin. These formulations are based on generalizations of the requirement of invariance under rotations in a three-dimensional isobaric space.

The first natural generalization is to retain the three-dimensional space but to consider the possibility of reflections. The consequent introduction of isobaric parity is due to Pais (1953) and the application of this idea to the scheme was made by Prentki and D'Espagnat (1956, see also Racah 1956, Salam 1956). The Λ^0 is considered as a true scalar and the Σ-multiplet and π-multiplet as pseudo-vectors in isobaric space. They are characterized therefore by an isobaric inversion quantum number $P_{\mathrm{I}} = +1$. That is, under inversion they all satisfy $\psi \rightarrow +\psi$. The K-mesons, nucleons and cascade particles are all spinors in isobaric space. Under reflection there are the familiar two types of spinor, the first and second kind that can be characterized by inversion quantum numbers $+i$ and $-i$ respectively. It is assumed that the K-mesons and nucleons are of the first kind and the $\bar{\mathrm{K}}$-mesons and the Ξ particles are of the second kind. If we define a quantity U by the relation

$$P_{\mathrm{I}} = (i)^U \qquad \ldots\ldots(7.23)$$

then the assignments of U for this scheme are summarized in table 5.

Table 5

$U = -1$	$U = 0$	$U = +1$
	Λ^0	
$\Xi^-, \bar{\Xi}^0$	$\Sigma^-, \Sigma^0, \Sigma+$	$\mathrm{N}^0, \mathrm{N}^+$
$\mathrm{K}^-, \bar{\mathrm{K}}^0$	π^-, π^0, π^+	$\bar{\mathrm{K}}^0, \mathrm{K}^+$

An ambiguity in the definition of U arises if isobaric pseudoscalar or vector particles exist. For these particles $P_{\mathrm{I}} = -1 = i^2 = i^{-2}$, and we cannot choose between $U = +2$ or $U = -2$. A certain choice of interaction schemes might lead us to choose one definition or the other, but the most general scheme does not.

Prentki and D'Espagnat therefore assume that U satisfies the condition

$$|U| \leqslant 1 \qquad \ldots\ldots(7.24)$$

which excludes isobaric pseudoscalars and vectors.

If the strong interactions are invariant under reflections in the isobaric space, then the product of P_1's is a constant of the motion. Moreover, if the basic interactions are all of the Yukawa type and only involve three particles, then the conservation of this product implies the conservation of U. Prentki and D'Espagnat define charge by the relation

$$Q = T_3 + \tfrac{1}{2}U. \qquad \ldots\ldots(7.25)$$

A comparison of (7.25) and (7.6) shows that

$$U = S + B, \qquad \ldots\ldots(7.26)$$

and so the conservation of U together with the known conservation of B implies the conservation of S.

An attractive feature of the scheme is its economy. If multiply charged particles are excluded, the rule (7.24) provides room for the description of all the known elementary particles and only leaves one unfilled gap. This is the isobaric scalar boson.† However, if this particle were sufficiently massive it would be extremely unstable with a lifetime of approximately 10^{-23} second.

An alternative generalization of isobaric space consists in considering rotations only but in a four-dimensional space [4].‡ This arises naturally because of the well-known connection between the rotation group in [4] and the rotation group in [3] discussed in §3. If $I_{\alpha\beta}(\alpha, \beta = 1, \ldots, 4)$ are the generators of the rotation group in [4] we may define T_i, μ_i by

$$\begin{aligned} T_i &= \tfrac{1}{2}(I_{4i} + I_{jk}), \\ \mu_i &= \tfrac{1}{2}(-I_{4i} + I_{jk}), \end{aligned} \qquad \ldots\ldots(7.27)$$

where (i, j, k) is a cyclic permutation of $(1, 2, 3)$. The finite representations of the rotation group in [4] may be labelled by the two quantum numbers (T', μ'), corresponding to the representations of the generators (7.27) of the two three-dimensional rotation groups. Those representations for which $T' + \mu'$ is integral are single-valued tensor representations, those for which $T' + \mu'$ is half-odd-integral are double-valued spinor representations.

Again, Pais (1954) was the first to use a four-dimensional isobaric space. However, his formalism included many particles whose counterparts in nature have not yet been found. A more economic scheme of this type is that proposed by Salam and Polkinghorne (1955) (SP). They propose that the nucleons and Ξ particles together form the $(\frac{1}{2}, \frac{1}{2})$ or vector representation, and that the K- and $\bar{\text{K}}$-mesons form another isobaric vector. The Σ particles and the pions are each a $(1, 0)$ representation, or self-dual antisymmetric tensor. The Λ is the $(0, 0)$ or scalar representation. The relation between charge and isobaric space is the simple one

$$\psi = I_{12} = T_3 + \mu_3. \qquad \ldots\ldots(7.28)$$

† Some interesting consequences of the existence of such a particle have been discussed by Baldin and Kabir (1958). The acid test is whether the reaction $\text{d} + \text{d} \rightarrow {}^4\text{He} + \sigma$ exists. Since d and ^{4}He have isobaric spin zero the particle σ in this reaction must also be an isobaric scalar.

‡ This space must be Euclidean (Higgs 1957).

The fact that Q is integral implies a restriction to single-valued representations only of the rotation group in [4]. A comparison of (7.6) and (7.28) shows that

$$\mu_3 = \tfrac{1}{2}(S+B). \qquad \ldots\ldots(7.29)$$

The assignments of μ_3 for the SP scheme are summarized in table 6.

Table 6

$\mu_3 = -\frac{1}{2}$	$\mu_3 = 0$	$\mu_3 = \frac{1}{2}$
	Λ^0	
Ξ^-, Ξ^0	Σ^-, Σ^0, Σ^+	N^0, N^+
K^-, $\bar{K}^0$	π^-, π^0, π^+	K^0, K^+

If the strong interactions are invariant under rotations in [4], then μ_3 is conserved and the known conservation of B implies the conservation of S. An immediate difficulty faced by this scheme, however, is why there should be so pronounced a splitting of the baryon vector $(\frac{1}{2}, \frac{1}{2})$ into the widely separated nucleon doublet ($\mu_3 = \frac{1}{2}$) and Ξ doublet ($\mu_3 = -\frac{1}{2}$). A possible resolution of this difficulty is discussed below.

The SP scheme has room for more particles than are at present known for certain in nature. In addition to the isobaric-scalar boson $(0, 0)$, there is room for both a fermion and a boson multiplet corresponding to the $(0, 1)$ representation, or anti-self-dual antisymmetric tensor. The charged members of those multiplets would be multiply strange. It may be that these representations will be required to explain some of the anomalous cosmic-ray events. There is some evidence (see Dalitz 1957 for a summary) for a heavy boson of mass about 1500 e.m. undergoing a cascade decay

$$\omega^+ \to K^+ + \pi^0 \to \mu^+ + \nu + \pi^0. \qquad \ldots\ldots(7.30)$$

In accordance with the rule (7.7) we may suppose that ω^+ is the doubly strange positively charged member of the $(0, 1)$ triplet. If the neutral member of this triplet were sufficiently massive it would be highly unstable under decay into pions since it would have zero strangeness.

7.3. *Global Symmetry*

Recent experiments on the photoproduction of K-mesons (Clegg *et al.* 1957, Donoho and Walker 1957, Silverman *et al.* 1957) have led to the postulation of further symmetry principles governing the strong interactions. The experiments seem to indicate that the coupling of K-mesons with baryons is an order of magnitude weaker than the coupling of pions with nucleons,

$$g_K^2 \sim 1, \quad g_\pi^2 \sim 15. \qquad \ldots\ldots(7.31)$$

Gell-Mann (1957 b) and Schwinger (1957) have accordingly suggested that the strong interactions may decompose into two classes: the medium strong (MS) or K-meson interactions, and the very strong (VS) or pion interactions. According to the currently accepted principle that the stronger the interaction the more extensive are its symmetry properties, one might expect that the MS interactions

would possess three-dimensional charge independence symmetry and that the VS interactions would possess further symmetries.

One way of introducing these additional symmetries would be to consider a four-dimensional isobaric space. In terms of the SP scheme one might suppose that only the VS interactions are invariant under the full four-dimensional rotation group (Polkinghorne 1957). The effect of this requirement is to couple the nucleons and the cascade particles to pions in an exactly similar manner. The MS interactions would then not be invariant under μ-rotations and would provide the explanation of the $N-\Xi$ mass difference.

The SP scheme uses the well-known group theoretical way of connecting a three-dimensional and a four-dimensional rotation group. Another way of effecting the connection is simply to consider rotations only in the three-space perpendicular to a given direction in the four-space. In terms of the first method a four-vector decomposes into two three-space spinors, while in terms of the second method it decomposes into a three-vector and a scalar (the fourth component). Schwinger (1957) supposes that nature employs both methods, the first giving N and Ξ, the second Σ and Λ. With respect to their pion interactions both groups are to behave like four-vectors and thus for their interactions N and Ξ behave similarly, and Σ and Λ behave similarly. Thus the pion–baryon coupling constants are reduced to two in number.

In fact, however, Gell-Mann and Schwinger both assume a greater degree of symmetry which they term *global symmetry*. This assumes that in the absence of MS and weaker interactions the baryons form a degenerate super-multiplet all of whose members have the same bare mass and pion interactions. The similarity of the pion interactions can be made precise in the following way (Gell-Mann 1957 b):

We define the four doublets

$$\mathrm{N}_1 = \begin{pmatrix} \mathrm{N}^+ \\ \mathrm{N}^0 \end{pmatrix}, \quad \mathrm{N}_2 = \begin{pmatrix} \Sigma^+ \\ (\Lambda^0 - \Sigma^0)/\sqrt{2} \end{pmatrix},$$

$$\mathrm{N}_3 = \begin{pmatrix} (\Lambda^0 + \Sigma^0)/\sqrt{2} \\ \Sigma^- \end{pmatrix}, \quad \mathrm{N}_4 = \begin{pmatrix} \Xi^0 \\ \Xi^- \end{pmatrix}. \qquad \ldots\ldots(7.32)$$

Then the pion–baryon interaction is assumed to be

$$ig\left[\sum_{i=1}^{4} \bar{\mathrm{N}}_i \tau\gamma_5 \mathrm{N}_i\right]\pi. \qquad \ldots\ldots(7.33)$$

The interaction (7.33) is identically equal to

$$ig[\bar{\mathrm{N}}\tau\gamma_5 \mathrm{N} + \bar{\Xi}\tau\gamma_5 \Xi + \bar{\Lambda}\gamma_5 \boldsymbol{\Sigma} + \bar{\boldsymbol{\Sigma}}\gamma_5 \Lambda + \bar{\boldsymbol{\Sigma}}_\Lambda \gamma_5 \boldsymbol{\Sigma}].\pi. \qquad \ldots\ldots(7.34)$$

Global symmetry produces a greater symmetry in pion–baryon interactions than is implied by any form of four-dimensional isobaric space. In 1952 Wigner predicted that all baryons would have the same coupling with pions. He argued that in this way the pion field would 'realize' the conservation of baryon number in a way similar to the 'realization' of the conservation of charge by the electromagnetic field. This attractive idea is, however, difficult to formulate in a precise form since the interaction with the electromagnetic field is through the conserved current j_μ, but this is not so for the pion interaction which takes place through a

non-conserved pseudoscalar quantity. The electromagnetic field can be introduced as an extension field (see § 4) associated with phase invariance, but the pion field cannot be introduced as the extension field connected with baryon number. Thus the relation, if any, of global invariance and baryon number remains obscure.

Schwinger's theory has a further invariance property. He assumes that the VS interactions involve not only the pseudoscalar isovector field $\boldsymbol{\pi}$ but also a scalar isoscalar field σ coupled by a term

$$g\left[\sum_{i=1}^{4}\bar{\mathrm{N}}_i\,\mathrm{N}_i\right]\sigma. \qquad \ldots\ldots(7.35)$$

The theory in the absence of all but the VS interactions is invariant under the infinitesimal transformations

$$\begin{aligned}\delta\mathrm{N}_i &= \tfrac{1}{2}i\gamma_5\,\delta\boldsymbol{\epsilon}.\boldsymbol{\tau}\mathrm{N}_i,\\ \delta\boldsymbol{\pi} &= -\delta\boldsymbol{\epsilon}\sigma,\\ \delta\sigma &= \delta\boldsymbol{\epsilon}.\boldsymbol{\pi},\end{aligned} \qquad \ldots\ldots(7.36)$$

except for the bare mass term in the free Lagrangian. However, this term,

$$m_0\sum_{i=1}^{4}\bar{\mathrm{N}}_i\,\mathrm{N}_i, \qquad \ldots\ldots(7.37)$$

clearly transforms like (7.35) and so we should just replace σ by $\sigma+m_0/g$. This interesting invariance property, involving simultaneous transformations in isobaric space and spin space, is interpreted by Schwinger as exhibiting the relation of $\boldsymbol{\pi}$ and σ as components of a four-vector.

Little can be said or suggested about the form of the MS interactions except that they must be sufficiently asymmetrical to produce the observed splitting of the baryons into four multiplets. If arguments depending upon mass splittings are thought to deal with a subject too mysterious for us today, Pais (1958 a) has obtained in another way a limit on the symmetry that can exist between the strong interactions of baryons. The two multiplets closest together are the Σ and Λ. If we neglect the effects depending on $(M_\Sigma-M_\Lambda)/M_\Lambda$, then the strong interaction Lagrangian that treats Σ and Λ symmetrically may be written as

$$\begin{aligned}&[ig_1\,\bar{\mathrm{N}}_1\gamma_5\boldsymbol{\tau}\mathrm{N}_1+ig_2(\bar{\mathrm{N}}_2\boldsymbol{\tau}\gamma_5\,\mathrm{N}_2+\bar{\mathrm{N}}_3\boldsymbol{\tau}\gamma_5\,\mathrm{N}_3)+ig_4\,\bar{\mathrm{N}}_4\boldsymbol{\tau}\gamma_5\,\mathrm{N}_4].\boldsymbol{\pi}\\ &\quad+if_1\sqrt{2}(\bar{\mathrm{N}}_1\gamma_5\,\mathrm{N}_2\mathrm{K}^0+\bar{\mathrm{N}}_1\gamma_5\,\mathrm{N}_3\mathrm{K}^+)+\text{h.c.}\\ &\quad+if_2\sqrt{2}(\bar{\mathrm{N}}_4\gamma_5\,\mathrm{N}_2\mathrm{K}^--\bar{\mathrm{N}}_4\gamma_5\,\mathrm{N}_3\bar{\mathrm{K}}^0)+\text{h.c.}\end{aligned} \qquad \ldots\ldots(7.38)$$

We must, of course, assume that Σ and Λ have the same parity with respect to the nucleon and we have in fact also assumed that the K is pseudoscalar though our results do not depend on this. The K-meson terms in (7.38) arise from rewriting

$$if_1(\bar{\mathrm{N}}\gamma_5\Lambda\mathrm{K}+\bar{\mathrm{N}}\boldsymbol{\tau}.\gamma_5\boldsymbol{\Sigma}\mathrm{K})+\text{h.c.}+if_2(\bar{\Xi}\gamma_5\Lambda\bar{\mathrm{K}}+\bar{\Xi}\boldsymbol{\tau}.\gamma_5\boldsymbol{\Sigma}\bar{\mathrm{K}})+\text{h.c.} \qquad \ldots\ldots(7.39)$$

The doublets K and $\bar{\mathrm{K}}$ are

$$\mathrm{K}=\begin{pmatrix}\mathrm{K}^+\\ \mathrm{K}^0\end{pmatrix},\quad \bar{\mathrm{K}}=\begin{pmatrix}-\bar{\mathrm{K}}^0\\ \mathrm{K}^-\end{pmatrix}. \qquad \ldots\ldots(7.40)$$

Many contradictions with experiment, both qualitative and quantitative can be extracted from (7.38). For example, it is invariant under the phase transformations :

$$\begin{aligned} &N_2 \to e^{i\alpha} N_2, \quad K^0 \to e^{-i\alpha} K^0, \quad \bar{K}^0 \to e^{i\alpha} \bar{K}^0, \\ &N_3 \to e^{i\beta} N_3, \quad K^+ \to e^{-i\beta} K^+, \quad K^- \to e^{i\beta} K^-, \\ &N_1 \to N_1, \quad N_4 \to e^{i(\alpha+\beta)} N_4. \end{aligned} \qquad \ldots\ldots(7.41)$$

This implies the separate conservation of the quantum numbers S_1 and S_2 whose assignments are given in table 7.

Table 7

	S_1	S_2
π, N	0	0
$K^{\pm}$	± 1	0
$K^0(\bar{K}^0)$	0	$+1(-1)$
N_2	0	-1
N_3	-1	0
Ξ	-1	-1

S_1 and S_2 are related to strangeness by

$$S = S_1 + S_2. \qquad \ldots\ldots(7.42)$$

Perhaps the most striking contradiction with experiment obtained from the separate conservation of S_1 and S_2 is that it forbids the charge exchange scattering of K-mesons

$$K^+ + N^0 \to K^0 + N^+. \qquad \ldots\ldots(7.43)$$

Since this is observed the strong interactions of Σ and Λ cannot have an identical form.

7.4. *Electric Charge*

The central mystery of all schemes is the connection between isobaric structure and electric charge (equations (7.6), (7.25) and (7.30)). In particular there is the problem of the weak interactions which do not conserve T_3 or S but do conserve Q. D'Espagnat, Prentki and Salam (1957) have recently made an interesting suggestion that links this question with the $|\Delta T| = \frac{1}{2}$ rule. In the four-dimensional SP formalism the charge is just I_{12} which together with I_{23}, I_{31} generates a three-dimensional rotation group, the group of rotations in the space perpendicular to the four-axis. D'Espagnat, Prentki and Salam suppose that the weak interactions involving strongly interacting particles are invariant under this rotation group (M-rotations). In M-space the Λ is a scalar, the Σ's and the π's three-vectors. The $N-\Xi$ quartet become a vector and a scalar,

$$\begin{aligned} \mathbf{N} &= \left(\frac{N^+ + \Xi^-}{\sqrt{2}}, \quad \frac{-N^+ + \Xi^-}{\sqrt{2}i}, \quad \frac{-N^0 + \Xi^0}{\sqrt{2}} \right), \\ N_S &= \frac{(\Xi^0 + N^0)}{\sqrt{2}}, \end{aligned} \qquad \ldots\ldots(7.44)$$

and the K-mesons decompose similarly. The terms (7.44) may be rewritten in the form (Prentki and D'Espagnat 1958)

$$\begin{aligned} \sqrt{2}\mathbf{N} &= x^* \boldsymbol{\tau} N + y^* \boldsymbol{\tau} \Xi, \\ \sqrt{2} N_S &= x^* N + y^* \Xi, \end{aligned} \qquad \ldots\ldots(7.45)$$

where $x^* = (0, 1)$ and $y^* = (1, 0)$ are 'spurions', that is numerical spinors. A term involving **N** or N_S once (or **K** or K_S once) will therefore correspond to a term giving $|\Delta T| = \frac{1}{2}$ carried away by the 'spurion'. The VS interactions are invariant under M-rotations and so any process occurring through a number of elementary weak and VS interactions only must have an equivalent single phenomenological interaction term that is a scalar in M-space. If the number of N, Ξ and K particles in both initial and final states is n, then clearly the interaction must satisfy $|\Delta T| \leqslant n/2$. Since the MS interactions are not invariant under M-rotations, their presence produces corrections to this rule.

D'Espagnat, Prentki and Salam extend their M-invariance to weak interactions involving leptons. Here, however, the theory faces many difficulties, not the least of which is that arising from combining the μ (mass 204 e.m.) in a multiplet with the electron and neutrino. It does not seem that this sort of invariance can be usefully applied to particles not possessing strong interactions.

§ 8. Conclusion

Something of the power of invariance principles in elementary particle physics has been shown above. The fact that at present these principles are so much in the centre of interest is due to two main reasons.

In the first place there are many problems, particularly those relating to strong interactions, in which we are mathematically powerless to give any detailed quantum field theoretical description of the phenomena and for which the main progress in recent years has come from the use of general principles such as invariance and causality. By means of these, a qualitative and sometimes even quantitative understanding of the phenomena has been achieved.

The second reason for the prominence of invariance considerations in present-day thinking is of course the unexpected discovery of the *violation* of at least one such principle, namely parity conservation.

The invariance principles we have considered fall into two classes. The first class consists of those whose origins are in some sense understood, for instance with the aid of the correspondence principle, when the invariance in question has meaning in classical physics. An example is invariance under the Lorentz group. This is a manifestation of the properties of the space–time continuum. We might expect, therefore, that any experimentally detected failure of this principle could be used to obtain more accurate information about the actual space–time structure. In practice it has not turned out as easy as that. In this class the only failure of invariance that has yet been detected concerns reflections, discrete transformations, whose properties are more recondite than those associated with continuous transformations, and this has in fact so far hardly given any greater depth to our understanding. It has, however, brought with it the new puzzle that violation of symmetry is *selective* in that it occurs only in the weak interactions.†

This feature is quite unexpected in the first class of invariance principles; at the same time it bridges the gap to the second class which consists of those mysterious invariances associated with isobaric spin and strangeness. We are

† An ingenious suggestion has been made by Feinberg (1957) and by Soloviev (1958). They suppose that the non-conservation of parity is potentially present, so to speak, in other interactions, but higher symmetries provide additional constraints that prevent its actually appearing.

ignorant of their origin or ultimate significance but we have always been very well aware of their violations, which appear to be of much the same nature as the selectiveness shown by the non-conservation of parity. Isobaric spin is not conserved by electromagnetic or weak interactions while the latter also do not conserve strangeness.

'The stronger the interaction, the greater the number of symmetries it will possess' has become a useful working maxim in physics. Although we have come to accept this we are far from understanding why it should be so.† Carried to its logical extreme it would imply that the gravitational interaction (which is so much weaker than those discussed here that it is of no practical account in elementary particle physics) would be completely chaotic. In fact there is evidence that this interaction has a highly complex group of associated symmetries.

Physics never ends with a full stop but only with a further question mark. This survey concludes with this problem of the selective failure of invariance principles.

APPENDIX

Form Invariance

In §2 we saw how invariance of the Hamiltonian of a system under a unitary transformation or a family of such transformations in general leads to a conservation law, and possibly to other consequences, and we have seen many examples of this in previous sections. Here we wish to point out that even if the Hamiltonian is not strictly invariant, but remains invariant in form, only with the values of some c-number parameters altered, valuable results may still be obtained, although of course there will be no conservation law.

The Hamiltonian H of the system will normally be a function of various field operators denoted generically by ϕ, and various c-number parameters denoted by c. In the original applications, the c-number parameters were beta-decay coupling constants. Similarly, states of the system will be labelled by the eigenvalues k, say, of a set of observables K which will be given by the theory as functions of the fields ϕ and parameters c. The physical meaning of the state will be determined via the correspondence principle, and will depend only on the physical interpretation of the observables K and the particular eigenvalues k, and not on the particular values of the parameters c used in the theory, or on any *a priori* interpretation of the fields ϕ which in general will not have any experimentally realizable correspondence limit. The particular vectors in Hilbert space which correspond to a definite set of eigenvalues k, when expressed as a linear combination of a set of base vectors in terms of which the fields ϕ are known operators with well-defined matrix elements, will, however, obviously depend on the particular values of the parameters c. In stating this, we merely give expression to the postulate that it is K, a function of both c and ϕ, rather than ϕ alone, which is of observational significance.

† An alternative attitude to this situation adopted by some (cf. Bludman 1958, also discussion between Peierls and Salam 1958) is that, possibly, weak and strong interactions are subject to conflicting sets of invariances, but that the full extent of the invariances governing the former is masked by the dominating strength of the latter.

We give expression to these concepts by writing

$$K = K(\phi, c), \qquad \ldots\ldots(A1)$$

$$H = H(\phi, c), \qquad \ldots\ldots(A2)$$

$$K(\phi, c)|k, c\rangle = k|k, c\rangle. \qquad \ldots\ldots(A3)$$

If we now consider a unitary transformation whereby all states $|\Psi\rangle$ are mapped on to $U|\Psi\rangle$ and all operators $\mathcal{O}$ are mapped on to $U\mathcal{O}U^{-1}$, then all matrix elements of the new operators between the new states are identical with the matrix elements of the corresponding old operators between the corresponding old states—nothing in the physical significance of the theory is altered by the transformation. We shall have

$$\phi' = U\phi U^{-1}, \qquad \ldots\ldots(A4)$$

$$K' = UKU^{-1} = K(\phi', c), \qquad \ldots\ldots(A5)$$

$$H' = UHU^{-1} = H(\phi', c), \qquad \ldots\ldots(A6)$$

$$|k, c\rangle' = U|k, c\rangle, \qquad \ldots\ldots(A7)$$

$$K'|k, c\rangle' = k|k, c\rangle'. \qquad \ldots\ldots(A8)$$

The theory is strictly invariant under transformation by U if $H' = H$ and $K' = K$. We shall call the theory *form-invariant* if we can find a new set of c-number parameters c' such that

$$H'(\phi, c) \equiv H(\phi', c) = H(\phi, c'), \qquad \ldots\ldots(A9)$$

$$K'(\phi, c) \equiv K(\phi', c) = K(\phi, c'), \qquad \ldots\ldots(A10)$$

that is, such that the unitary transformation is equivalent to a change in the values of the parameters of the theory. Then it follows from (A8) that

$$K(\phi, c')|k, c\rangle' = k|k, c\rangle', \qquad \ldots\ldots(A11)$$

so that, apart from a phase-factor,†

$$U|k, c\rangle \equiv |k, c\rangle' = |k, c'\rangle.$$

The phase factor may be written $\exp[i\delta(k, c, c')]$.

Now, the matrix element of any operator $A(\phi, c)$, which is also form-invariant so that $UAU^{-1} = A(\phi, c')$, between eigenstates of K with eigenvalues k_1 and k_2 will be

$$\begin{aligned} a(k_1, k_2, c) &\equiv \langle k_2, c|A(\phi, c)|k_1, c\rangle, \\ &= \langle k_2, c|U^{-1}UA(\phi, c)U^{-1}U|k_1, c\rangle, \\ &= \langle k_2, c'|A(\phi, c')|k_1, c'\rangle \exp\{i[\delta(k_1, c, c') - \delta(k_2, c, c')]\}, \\ &= a(k_1, k_2, c') \exp\{i[\delta(k_1, c, c') - \delta(k_2, c, c')]\}, \qquad \ldots\ldots(A12) \end{aligned}$$

† We assume here that the physical state is defined uniquely by the eigenvalues k, i.e. that there is no degeneracy. The extension to the degenerate case will be indicated later.

so that all such matrix elements can depend on the parameters c only in a manner which, apart from a phase factor, is invariant under the substitution $c \to c'$. This will be true in particular for matrix elements of the S-matrix, since this is constructed from H which by hypothesis is form-invariant. In most applications, the phase factor either can be taken to be 1 by suitable definition of $|k, c\rangle$, or else is obvious as, for example, where U merely changes the relative phases of states containing different numbers of particles.

This type of form invariance of the observable consequences of a theory under a unitary transformation leaving the operators representing observables form-invariant, which is expressed by (A12), was first stated by Pauli (1957) and independently by Pursey (1957). It can be extended to situations where it is impossible to define a complete set of form-invariant operators K for the purpose of defining initial and final states, but where one is only interested in averages of the matrix elements of operators, or of the squared modulus of such matrix elements, over sets of initial and final states where each set of states can be defined as a (degenerate) set of eigenstates of form-invariant operators (Kahana and Pursey 1957, Ferretti 1957).

The principle described has so far been applied mainly to the analysis of the dependence of weak decay processes on the coupling constants for very general interactions. In addition to the papers quoted, the chief work has been done by Lüders (1958) and Gatto and Lüders (1958). Although the need to consider the most general possible interaction in such processes seems to have passed, we have felt it worth exploring the principle for the sake of possible new applications in a different context.

The form invariance principle also gives deeper insight into the well-known independence of the predictions of quantum field theory of the particular representation chosen for the Dirac matrices. The usual argument is, of course, that when detailed calculations are made it is found that the Dirac matrices enter only through traces of products, which are representation-invariant, and this can be seen to be generally true on making an analysis of the nature of calculations with the theory. The form invariance argument does not require any such detailed analysis of the computational method, but rests instead on the general principles of quantum mechanical transformation theory.

In field theory these matrices are not operators but c-numbers—the matrix notation is just a convenient notation to express various linear and bilinear combinations of field operators.† All operators of physical significance, however, such as the Hamiltonian, total momentum, angular momentum, number operators, etc., are form-invariant under those unitary transformations of the theory which are equivalent to a change of representation of the Dirac matrices. Hence all matrix elements of all physically significant operators between states defined as eigenstates of physically significant observables are independent of the representation of the Dirac matrices. They do, of course, depend on those combinations of the matrix elements of the Dirac matrices which are invariant under change of representation : these are just the matrix elements of the anticommutation rules.

† More precisely, we regard the matrix elements of the γ-matrices as c-number parameters which can be modified by a unitary transformation of the quantized field theory in such a way as to preserve the relations between them which are summarized in the anticommutation rules for the Dirac matrices.

Another trivial consequence of form invariance is Nishijima's (1957) result that matrix elements between physically significant states of the S-matrix in electrodynamics are invariant under the transformation $\psi \to \exp(i\theta\gamma_5)\psi$ of the electron field. This may be seen either as a change of representation of the Dirac matrices, or else by introducing an additional parameter α into the theory by writing the mass term as $m\exp(i\alpha\gamma_5)$, which becomes $m\exp\{i(\alpha+2\theta)\gamma_5\}$ when ψ is replaced by $\psi\exp(i\theta\gamma_5)$, and using form invariance under the transformation considered to show that S-matrix elements are independent of α.

The principle of form invariance can be further extended to the case where some of the observables K undergo a linear transformation, so that the eigenvalues labelling the transformed states are different from those labelling the original state :

$$U|k, c\rangle = |k', c'\rangle \times \text{phase factor.} \qquad \ldots\ldots\text{(A13)}$$

In this case, matrix elements of form-invariant operators must be, apart from a phase factor, a function of k, c which is invariant under the substitutions $k \to k'$, $c \to c'$. Still further generalizations in the same spirit are possible.

References

Baldin, A. M., and Kabir, P. K., 1958, *Nuovo Cim.*, **9**, 547.

Bell, J. S., 1955, *Proc. Roy. Soc.* A, **231**, 479.

Biedenharn, L. C., and Rose, M. E., 1951, *Phys. Rev.*, **83**, 459 ; 1953, *Rev. Mod. Phys.*, **25**, 729.

Blatt, J. M., and Weisskopf, V. F., 1952, *Theoretical Nuclear Physics* (New York : John Wiley).

Bludman, S. A., 1958, *Nuovo Cim.*, **9**, 433.

Breit, G., Condon, E. U., and Present, R. D., 1936, *Phys. Rev.*, **50**, 825.

Burcham, W. E., 1955, *Progress in Nuclear Physics*, **5**, 171 (London : Pergamon Press).

Cassen, B., and Condon, E. U., 1936, *Phys. Rev.*, **50**, 846.

CERN, 1958, *Proceedings of the CERN Conference on High Energy Nuclear Physics, Geneva*, 1958 (Geneva : CERN).

Clegg, A. B., Ernstene, M. P., and Tollestrup, A. V., 1957, *Phys. Rev.*, **107**, 1200.

Coester, F., 1951, *Phys. Rev.*, **84**, 1259.

Dalitz, R. H., 1957, *Rep. Progr. Phys.*, **20**, 163 (London : Physical Society).

D'Espagnat, B., Prentki, J., and Salam, A., 1957, *Nuclear Physics*, **5**, 447.

Donoho, P. L., and Walker, R. L., 1957, *Phys. Rev.*, **107**, 1198.

Edmonds, A. R., 1957, *Angular Momentum in Quantum Mechanics* (Princeton : Princeton University Press).

Feenberg, E., and Pake, G. E., 1953, *Notes on the Quantum Theory of Angular Momentum* (Cambridge, Mass. : Addison Wesley).

Feinberg, G., 1957 *Phys. Rev.*, **108**, 878.

Feldman, D., 1956, *Phys. Rev.*, **103**, 254.

Ferretti, B., 1957, *International Conference on Mesons and Recently Discovered Particles, Venice and Padua*, 1957, to be published in *Suppl. Nuovo Cim.*

Fröhlich, H., Heitler, W., and Kemmer, N., 1938, *Proc. Roy. Soc.* A, **166**, 154.

Gatto, R., and Lüders, G., 1958, *Nuovo Cim.*, **7**, 806.

Gell-Mann, M., 1953, *Phys. Rev.*, **92**, 833 ; 1957 a, *Nuovo Cim.*, **5**, 758 ; 1957 b, *Phys. Rev.*, **106**, 1296.

Gell-Mann, M., and Pais, A., 1955 a, *Proceedings of the* 1954 *Glasgow Conference on Nuclear and Meson Physics* (London : Pergamon Press) ; 1955 b, *Phys. Rev.*, **97**, 1387.

Gell-Mann, M., and Rosenfeld, A. H., 1957, *Annual Review of Nuclear Science*, **7**, 407.

Gürsey, F., 1958, *Nuovo Cim.*, **7**, 411.

Harish-Chandra, 1946, *Proc. Roy. Soc.* A, **186**, 502.

HEISENBERG, W., 1932, *Z. Phys.*, **77**, 1.
HIGGS, P. W., 1957, *Nuclear Physics*, **4**, 221.
JAUCH, J. M., and ROHRLICH, F., 1955, *The Theory of Photons and Electrons* (Cambridge, Mass. : Addison Wesley).
KAHANA, S., and PURSEY, D. L., 1957, *Nuovo Cim.*, **6**, 1469.
KEMMER, N., 1938 a, *Proc. Roy. Soc.* A, **166**, 127 ; 1938 b, *Proc. Camb. Phil. Soc.*, **34**, 354.
LANDAU, L., 1957, *Nuclear Physics*, **3**, 127.
LANDÉ, K., BOOTH, E. T., IMPEDUGLIA, J., LEDERMAN, L. M., and CHINOWSKI, W., 1956, *Phys. Rev.*, **103**, 1901.
LEE, T. D., 1955, *Phys. Rev.*, **99**, 337.
LEE, T. D., OEHME, R., and YANG, C. N., 1957, *Phys. Rev.*, **107**, 340.
LEE, T. D., and YANG, C. N., 1956, *Phys. Rev.*, **104**, 254 ; 1957, *Ibid.*, **105**, 1671.
LÜDERS, G., 1954, *K. Danske Vidensk. Selsk., Mat.-Fys. Medd.*, **28**, No. 5 ; 1957, *Annals of Physics*, **2**, 1 ; 1958, *Nuovo Cim.*, **7**, 171.
MAJORANA, M., 1937, *Nuovo Cim.*, **14**, 171.
MORPURGO, G., and TOUSCHEK, B. F., 1955, *Nuovo Cim.*, **1**, 1159.
MORPURGO, G., TOUSCHEK, B. F., and RADICATI, L. A., 1954, *Nuovo Cim.*, **12**, 677.
NISHIJIMA, K., 1954, *Progr. Theor. Phys., Japan*, **12**, 107, **13**, 285 ; 1957, *Nuovo Cim.*, **5**, 1349.
PAIS, A., 1952, *Phys. Rev.*, **86**, 633 ; 1953, *Physica*, **19**, 869 ; 1954, *Proc. Nat. Acad. Sci., Wash.*, **40**, 484 ; 1958 a, *Phys. Rev.*, **110**, 574 ; 1958 b, *Ibid.*, **112**, 624.
PAIS, A., and PICCIONI, O., 1955, *Phys. Rev.*, **100**, 1487.
PAIS, A., and TREIMAN, S., 1958, *Phys. Rev.*, **109**, 1759.
PAULI, W., 1955, Article in *Niels Bohr and the Development of Physics* (London : Pergamon Press) ; 1957, *Nuovo Cim.*, **6**, 204.
PEIERLS, R. E., and SALAM, A., 1958, *Proc. Roy. Soc.* A, **246**, 487.
PLANO, R., SAMIOS, N., SCHWARTZ. M., and STEINBERGER, J., 1957, *Nuovo Cim.*, **5**, 216.
POLKINGHORNE, J. C., 1957, *Nuovo Cim.*, **6**, 864.
PRENTKI, J., and D'ESPAGNAT, B., 1956, *Nuclear Physics*, **1**, 33 ; 1958, *Ibid.*, **6**, 596.
PURSEY, D. L., 1957, *Nuovo Cim.*, **6**, 266.
RACAH, G., 1956, *Nuclear Physics*, **1**, 302.
ROSE, M. E., 1957, *Elementary Theory of Angular Momentum* (New York : John Wiley).
SALAM, A., 1956, *Nuclear Physics*, **2**, 173 ; 1957, *Nuovo Cim.*, **5**, 299.
SALAM, A., and POLKINGHORNE, J. C., 1955, *Nuovo Cim.*, **4**, 848.
SCHWINGER, J., 1951, *Phys. Rev.*, **82**, 914 ; 1957, *Annals of Physics*, **2**, 407.
SEGRÈ, E., 1953, *Experimental Nuclear Physics* (New York : John Wiley).
SERPE, J., 1952, *Physica*, **18**, 295.
SHAW, R., 1955, *Ph.D. Dissertation*, Cambridge University.
SILVERMAN, A., WILSON, R. R., and WOODWARD, W. M., 1957, *Phys. Rev.*, **108**, 501.
SOLOVIEV, V. G., 1958, *Nuclear Physics*, **6**, 618.
SORRELS, J. D., LEIGHTON, R. B., and ANDERSON, C. D., 1955, *Phys. Rev.*, **100**, 1457.
TOLHOEK, H. A., and DE GROOT, S. R., 1951, *Phys. Rev.*, **84**, 151.
TOUSCHEK, B. F., 1957, *Nuovo Cim.*, **6**, 1281.
WATANABE, S., 1951, *Phys. Rev.*, **84**, 1008.
WICK, G. C., WIGHTMAN, A. S., and WIGNER, E. P., 1952, *Phys. Rev.*, **88**, 101.
WIGHTMAN, A. S., and SCHWEBER, S. S., 1955, *Phys. Rev.*, **98**, 812.
WIGNER, E. P., 1932, *Göttinger Nachr.*, 546 ; 1952, *Proc. Nat. Acad. Sci., Wash.*, **38**, 449.
WU, C. S., AMBLER, E., HAYWARD, R. W., HOPPES, D. D., and HUDSON, R. P., 1957, *Phys. Rev.*, **105**, 1413.
YANG, C. N., and TIOMNO, J., 1950, *Phys. Rev.*, **79**, 495.
YUKAWA, H., and SAKATA, S., 1937, *Proc. Phys.-Math. Soc., Jap.*, **19**, 1084.
YUKAWA, H., SAKATA, S., KOBAYASI, M., and TAKETANI, M., 1938, *Proc. Phys.-Math. Soc., Jap.*, **20**, 720.
YUKAWA, H., SAKATA, S., and TAKETANI, M., 1938, *Proc. Phys.-Math. Soc., Jap.*, **20**, 319.

NUCLEAR MAGNETISM IN PURE LIQUIDS

By J. G. POWLES
Physics Department, Queen Mary College (University of London)

CONTENTS

	PAGE
§ 1. Introduction	433
§ 2. Magnetic shielding effects in steady state spectra	439
2.1. Introduction	439
2.2. Magnetic shielding in atoms	441
2.3. Magnetic shielding in molecules	442
2.3.1. Perturbation methods	442
2.3.2. Variational methods	446
2.3.3. Circulating currents	449
2.3.4. Intermolecular effects	452
2.3.5. Intramolecular effects	455
2.3.6. Anisotropy relaxation	457
2.3.7. Chemical structure and analysis	458
§ 3. Indirect spin–spin interactions in steady state spectra	460
3.1. Introduction	460
3.2. Phenomenological explanation	463
3.3. The physical origin of indirect spin–spin coupling	470
3.4. Missing multiplets	477
3.5. Double resonance in high resolution spectra	478
3.6. Analysis of complex spectra	481
3.7. Multiple quantum transitions	484
3.8. Natural line widths	488
§ 4. Magnetic shielding and indirect spin–spin interaction effects in transient experiments	489
4.1. Introduction	489
4.2. Decay signals	489
4.3. Echoes	494
References	499

Abstract. This article is concerned with nuclear magnetic resonance effects in pure liquids which result directly or indirectly from the interactions of electrons with nuclei in diamagnetic molecules. Magnetic shielding and indirect spin–spin coupling are discussed in detail both experimentally and theoretically for steady state spectra and for transient measurements. Experimental techniques as such are not treated. The structure and behaviour of liquids is only dealt with in so far as it affects the interactions referred to above. The emphasis is placed on the description and explanation of physical effects rather than any detailed survey of their application to chemical and specifically molecular structure problems. There are some two hundred and fifty references and it is hoped that all the relevant ones have been included up to those available in London at August 1958.

§ 1. Introduction

The subject of magnetic resonance has become so vast that a comprehensive review of it would fill several volumes. This review is restricted, somewhat arbitrarily, to the nuclear magnetic induction effects in pure mobile liquids. The object of this restriction is to be able to discuss for nuclear magnetism some effects of molecular motion in liquids and the molecular effects due to electrons,

the magnetic shielding, the so-called chemical shift and the indirect spin–spin interaction between nuclei. The restriction will not be observed if it is advantageous to go outside it.

The basic principles of nuclear magnetic resonance are discussed very adequately in many reviews and books (see General References), and we give only a brief introduction here.

A nucleus of magnetic moment $\boldsymbol{\mu}$ in a steady magnetic field $\mathbf{H}_0$ experiences a torque $\boldsymbol{\mu} \wedge \mathbf{H}_0$ which as a consequence of its spin angular momentum $I\hbar$ causes it to precess about the field direction at the Larmor frequency ν_r where

$$2\pi\nu_r = \mu H_0/I\hbar = \gamma H_0. \qquad \ldots\ldots(1)$$

γ is called the gyromagnetic ratio and is characteristic of the nucleus. If a second magnetic field H_1 is applied at right angles to H_0 it will tend to cause the nucleus to precess (or nutate) about itself and so change the energy of the nucleus with respect to the field H_0. However the effect will be quite small (when $H_1 \ll H_0$) unless the field H_1 is rotating about the axis of H_0 at or near the Larmor frequency ν_r, since otherwise the torque tends to average to zero. We therefore have a resonant effect.

Alternatively one may say that the nucleus has only certain allowed orientations in the field H_0 and so a number of discrete energy levels depending on its spin.

The Hamiltonian for a nucleus in a magnetic field $\mathbf{H}$ is

$$\mathscr{H}_m = -\boldsymbol{\mu}.\mathbf{H} \qquad \ldots\ldots(2)$$

where $\boldsymbol{\mu}$ is the magnetic moment vector which is parallel to $\mathbf{I}$ the spin vector and according to the definition of γ in (1),

$$\boldsymbol{\mu} = \hbar\gamma\mathbf{I} \qquad \ldots\ldots(3)$$

so that

$$\mathscr{H}_m = -\hbar\gamma\mathbf{I}.\mathbf{H}. \qquad \ldots\ldots(4)$$

For a static magnetic field H_0 in the z direction $\mathbf{H} = (0, 0, H_0)$ and

$$\mathscr{H}_m = -\hbar\gamma I_z H_0. \qquad \ldots\ldots(4')$$

The eigenfunctions are ϕ_m where m has the $(2I+1)$ values $-I, -I+1, \ldots I$ and $I_z\phi_m = m\phi_m$; thus if $\mathscr{H}_m\phi_m = E_m\phi_m$, $E_m = -\hbar\gamma H_0 m$ and the levels are uniformly spaced with intervals $\Delta E = \hbar\gamma H_0$. For magnetic dipole interaction with radiation the selection rule is $\Delta m = \pm 1$ and so the absorbed, or emitted, quanta of radiation are of energy $\hbar\gamma H_0$ which must equal $h\nu_r$, i.e. $2\pi\nu_r = \gamma H_0$ which is the same as the classical result equation (1). For convenience we shall consider a nucleus with $I = \frac{1}{2}$ and if $I > \frac{1}{2}$ we shall make specific mention of any modifications required. If $I = \frac{1}{2}$ we put $\phi_{\frac{1}{2}} = \alpha$ and $\phi_{-\frac{1}{2}} = \beta$ and use the well-known relations

$$I_z\binom{\alpha}{\beta} = \tfrac{1}{2}\binom{\alpha}{-\beta}, \quad I_+\binom{\alpha}{\beta} = \binom{0}{\alpha}, \quad I_-\binom{\alpha}{\beta} = \binom{\beta}{0} \quad \text{where} \quad I_\pm = I_x \pm iI_y. \qquad \ldots\ldots(5)$$

(For any I, $I_\pm\phi_m = [I(I+1) - m(m+1)]^{\frac{1}{2}}\phi_{m\pm1}$.)

The time dependent wave functions are then $\alpha\exp(\frac{1}{2}.2\pi i\nu_r t)$ and $\beta\exp(-\frac{1}{2}.2\pi i\nu_r t)$ and the complete wave function when $\mathscr{H}_m$ is the largest term in the Hamiltonian may be written $\psi(t) = a(t)\,\alpha\exp(\frac{1}{2}.2\pi i\nu_r t) + b(t)\,\beta\exp(-\frac{1}{2}.2\pi i\nu_r t)$

where the populations in the levels α and β are aa^* and bb^*. The time dependence of $a(t)$ and $b(t)$ indicate transitions between the levels. In order to change the energy of the nucleus in the field H_0 we must supply radiation of the correct frequency and polarization to cause transition and so get a resonant effect or a spectroscopic

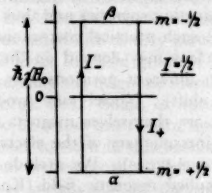


Figure 1.

absorption line. For an applied rotating H_1 field $(H_1 \cos 2\pi\nu t, -H_1 \sin 2\pi\nu t, 0)$ there is a term in the Hamiltonian

$$\mathscr{H}_1 = -\gamma\hbar\mathbf{I}.\mathbf{H}_1 = -\gamma\hbar H_1(I_x \cos 2\pi\nu t - I_y \sin 2\pi\nu t)$$
$$= -\tfrac{1}{2}\hbar\nu_1\{I_+ \exp(2\pi i\nu t) + I_- \exp(-2\pi i\nu t)\}$$

where $2\pi\nu_1 = \gamma H_1$ and ν_1 is the nutation frequency. $\mathscr{H}_1$ in first approximation (when $H_1 \ll H_0$) causes transitions between the levels which depend on the matrix elements $\langle m'|\mathscr{H}_1|m\rangle$ which are only non-zero when $m' = m \pm 1$, the selection rule, and the probability oscillates violently unless $\nu \simeq \nu_r$, the resonance condition. The magnetization in the transverse direction, which produces the signal, is proportional to $\langle \nu^*(t)|I_{x,y}|\nu(t)\rangle$.

A valuable aid to understanding these phenomena is to use rotating coordinate systems, or their equivalent in quantum mechanics. For an applied field $(H_1 \cos 2\pi\nu t, -H_1 \sin 2\pi\nu t, H_0)$ the effective field in the coordinate system rotating about Oz at ν is $\mathbf{H}_r = (H_1, 0, H_0 - 2\pi\nu/\gamma)$ which removes the time dependence of the magnetic field and reduces the complex motion to a simple precession about $\mathbf{H}_r$ in the absence of relaxation. The corresponding procedure in the quantum mechanical case is a similarity transformation by the operator $\exp(2\pi i\nu I_z t)$. The Hamiltonian $-\gamma\hbar\mathbf{I}.\mathbf{H}$ becomes just $-\gamma\hbar\mathbf{I}.\mathbf{H}_r$ in the rotating system.

In practice the resonant absorption line is not infinitely sharp or even necessarily single and the line broadening may be due to many causes some of which will be discussed in detail in this review. The simplest cause of line broadening is the inhomogeneity of the field H_0 in which case different nuclei are in slightly different fields and therefore have slightly different precession frequencies, or energy level separations. This may be represented by adding a term $-\gamma\hbar I_z \Delta H_0$ to the Hamiltonian with the appropriate distribution over ΔH_0. Consequently the absorption takes place over a range of frequency or of field, whichever is varied. It is of paramount importance to note that the field which is significant is that *at the nucleus* and that this may be different from the applied field due to various and subtle effects. We restrict ourselves in this review to situations in

which the material is not paramagnetic, e.g. molecules in totally symmetric ground states for which the electrons have strictly zero orbital and spin angular momentum. In this case the magnetic field due to the electrons at the nuclei is zero in the free molecule and we are thus enabled to study more subtle effects.

The magnetic field inside the specimen, subjected to an external field, depends on its shape and its bulk magnetic properties and thus so does that at the nucleus. We shall usually assume that such practical corrections have been made.

The actual field at a nucleus may depend on the position of the nucleus in the molecule since nuclei in different positions may have a different electronic environment (the chemical shift). Nuclei may produce fields at other nuclei either directly because they are themselves magnets (direct dipole–dipole interaction) or indirectly by the intermediary of the electrons in the system (indirect spin–spin interactions of various types). We exclude effects due to free electron spin interaction with the applied magnetic field (Knight shift, effects in ferromagnetic materials, etc.) which are negligible in the cases under discussion.

The resonance line may also be affected, if the nucleus has a spin $I > \frac{1}{2}$ by the interaction of its electric quadrupole moment Q with an electric field gradient which is usually a field 'built in' the molecule. This provides an additional torque on the nucleus which may be regarded as affecting its precessional motion in the magnetic field or as giving additional terms in the Hamiltonian and so modifying the energy levels. (The additional terms in the Hamiltonian are of the form, $-\mathbf{Q}.\nabla\mathbf{E}$ where $\mathbf{Q}$ and $\nabla\mathbf{E}$ are the nuclear electric quadrupole and the field gradient tensor operators, the first of which may be expressed in terms of $\mathbf{I}$ so that the formalism already described can be used in calculations.) Since we shall be mainly concerned with nuclei of spin $\frac{1}{2}$ and in any event most quadrupole effects are averaged out, this will not be discussed in any detail.

There are numerous effects due to motion of the nuclei. If the nuclei move about rapidly enough in the inhomogeneous magnetic field, either by diffusion or by actual macroscopic motion of the sample, they may experience only an average field and the line may be narrowed. Macroscopic motion is in fact used to reduce residual inhomogeneity of the magnetic field H_0 (sample spinning).

The sources of line broadening or complexity are sensitive to motion in varying degree. The direct dipole–dipole interaction may be averaged out to any required extent by sufficiently extensive and rapid variation of the orientation and length of the line joining two interacting nuclei. The idea of motional or collision *narrowing* as opposed to the collision broadening familiar in other parts of spectroscopy merits explanation. In ordinary spectroscopy the time between collisions is longer than the period of oscillation and every collision produces a transition; so we have the broadened spectrum of an interrupted harmonic oscillator. In nuclear resonance in the static case (solid) there is a constant perturbation which broadens the energy levels and therefore the absorption line. In a liquid the perturbation varies rapidly and must be averaged. It may then be regarded as a series of weak collisions with time interval of the order 10^{-11} sec which is much less than the precession period ($\sim 10^{-7}$ sec). Thus the average effect over a period is small and the oscillation is not interrupted.

On the other hand the magnetic shielding or chemical shift effect tends to remain, although modified, in that, for instance, the effects due to anisotropy of the

shielding will tend to be removed. Any effect independent of orientation tends to remain.

It may happen also that nuclei exchange positions and if the local field is different at the two positions then a sufficiently rapid exchange will result in the two nuclei seeing not two different fields respectively but an average field and so the absorption line will be modified (nuclear exchange effects). For nuclei with $I > \frac{1}{2}$ electric quadrupole interaction effects may play a role. This however tends to be a small one in liquids since the field gradient $\nabla\mathbf{E}$ which tends to align the quadrupole moment (and therefore the magnetic moment) moves with the molecule and its effects tend to average out. Some exceptions will however be noted. Quadrupole effects may also be important for the spin–lattice relaxation time T_1, which is not however of direct concern in this article. The interactions between nuclei may also be changed by externally applied fields. The structure of a resonance line of one nuclear species may be due, at least in part, to interaction with another nuclear species. This interaction can be modified by sufficiently powerful irradiation of the second set of nuclei at its own resonant frequency (double resonance experiments). The observed signal can be very materially changed according to the intensity of that radiation.

There are saturation effects which arise from a disturbance of the populations of the energy levels and from a modification of the relaxation processes owing to the presence of the radiation. The line widths may be modified by radiation due to lifetime limitation.

For sufficiently intense irradiation the selection rules which apply for vanishing weak irradiation break down (multiple quantum transitions).

Another property which plays an important part in saturation experiments is the spin–lattice relaxation time but we shall not discuss this except in so far as it influences line shapes.

We have so far discussed the line shape observed in a steady state absorption experiment (or by observing an induction signal, which is essentially the same). It is also possible to carry out transient experiments in which the effects mentioned above play a role. If we produce a nuclear magnetization transverse to the main field H_0 we can observe it as a function of time by measuring the signal it induces in a coil whose axis is at right angles to H_0. In the simplest case the signal decays exponentially with a time constant T_2, the spin–spin or transverse relaxation time. T_2 is related to the line width δH in the steady state experiment by

$$\gamma\delta H = \alpha/T_2 \qquad \ldots\ldots(6)$$

where α is a constant of order unity. Since the transverse magnetization $M_{x,y}$ is $\sum_i \mu_{ix,y}$, T_2 is the time for the precessing nuclear magnets to get appreciably out of phase. For two nuclei in fields differing by δH the difference in precession frequency is $(\gamma/2\pi)\delta H$ and the time for them to get one radian out of phase is $1/\gamma\delta H$ which is approximately T_2, hence equation (6). When the line is complex the behaviour of the transverse magnetization with time is complex, but it is closely related to the Fourier transform of the line shape.

In a transient experiment the nuclear magnetization in the direction of the main magnetic field may not have its equilibrium value and if the system is left to itself equilibrium will often be restored exponentially with time constant T_1.

T_1 is called the spin–lattice or longitudinal relaxation time because the process involves exchange of energy between the spin system and the thermal bath, conventionally called the lattice even in a liquid or gas. Since the longitudinal magnetization is determined only by the population of the nuclei in the available energy levels and not by the phase of the nuclei in their precessional motion the longitudinal relaxation tends to be much simpler in behaviour than the transverse. Evidently T_1 and T_2 are different parameters since T_2 effects (dephasing) do not necessarily involve exchange of energy of a nucleus with the lattice but usually exchange between nuclei. However since exchange of energy of a nucleus with a lattice also causes loss of phase memory the difference in nature of T_1 and T_2 is not always very great. Indeed, in liquids with which we are primarily concerned we have $T_2 \simeq T_1$ in the simplest case. This indeed must be so when the ' collision time ' is much less than the nuclear Larmor frequency since if we transform to a coordinate system rotating at the Larmor frequency the interaction during the collision is almost unchanged and the nucleus is now in an isotropic situation, the field H_0 having been effectively removed. There is therefore, for the nucleus, no longer any distinction between the longitudinal and transverse directions, i.e. $T_1 = T_2$.

The transient experiments are more sensitive to some effects than others. A particular advantage of the spin echo technique is that it is possible to circumvent the effects of the inhomogeneity of the main field H_0 which is the most difficult technical problem in high resolution magnetic resonance. Indeed the highest resolution may still be obtained by the transient method since there appears to be no serious experimental limitation (except perhaps due to self-diffusion) here as there will always be in the production of magnetic fields. Transient experiments are sensitive to effects which are not measurable in the steady state experiment, e.g. self-diffusion effects and other time dependent properties which need not be averaged out. The two methods are complementary to each other.

The essentially irreversible processes associated with magnetic relaxation, which is an involved problem in non-equilibrium quantum statistical mechanics, may be elegantly dealt with by the density matrix method. If ρ is the density matrix we have for the Hamiltonian $\mathscr{H}$, $i\hbar\dot{\rho} = [\mathscr{H}, \rho]$. The assumption is used that the lattice is always in equilibrium. The expectation value of an operator Q is $\mathrm{Tr}\, Q\rho$.

We shall attempt no more than a brief description of the experimental techniques. The major problem in the high resolution experiments is the manufacture of a field H_0 of sufficient homogeneity and stability (say, 1 in 10^8). Several suitable magnets, and indeed complete high resolution equipments, are now available commercially and it would be invidious to make comparisons. It seems to this reviewer that the manufacture of suitable fields is much easier than was at first supposed and that the cost should fall considerably in the near future. This change has been brought about by an accretion of new materials and new ideas and perhaps principally by a sympathetic cooperation between engineers and physicists which is essential even in this field of ' light ' physics. It will not be long before 1 in 10^8 is as common in this field as, say, in the measurement of frequency, and the search for smaller and smaller interactions is likely to be fruitful. In some beam magnetic resonance experiments the sensitivity to small interactions is such as to make a worth-while preliminary test of the possible effect of purely nuclear forces at ångström distances !

The field under review may be expected to contribute a great deal to our knowledge of molecular structure as well as of liquids and already a significant contribution has been made. The great power of radio-frequency spectroscopy is its ability to detect small energy differences which are nevertheless of great importance since they may reflect very important properties of the system.

§2. Magnetic Shielding Effects in Steady State Spectra

2.1. *Introduction*

In diamagnetic materials the main broadening of a magnetic resonance line would be expected to be due to direct nuclear magnetic dipolar interaction. The classical dipolar interaction energy of two magnets at a separation $\mathbf{r}_{ij}$ is

$$\frac{\boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_j}{r_{ij}^3} - 3\frac{(\boldsymbol{\mu}_i \cdot \mathbf{r}_{ij})(\boldsymbol{\mu}_j \cdot \mathbf{r}_{ij})}{r_{ij}^5} \qquad \ldots\ldots(7)$$

which is readily converted to an operator by equation (3). The correctness of this interaction has been abundantly confirmed by measurements on solids (Pake 1958 and later workers). In the static case the effect is a splitting of the line due to the secular part

$$\mathscr{H}_{\text{dsec}} = \sum \gamma^2 \hbar^2 r_{ij}^{-3}(1-3\cos^2\theta_{ij})[I_{zi}I_{zj} - \tfrac{1}{4}(I_{-i}I_{+j} + I_{+i}I_{-j})] \qquad \ldots\ldots(8)$$

where θ_{ij} is the angle of r_{ij} with respect to H_0. This term is readily dealt with since $\mathscr{H}_d$ may usually be treated as a small perturbation of the Zeeman levels due to $\mathscr{H}_m$. The actual splitting may however be appreciable and line widths of several tens of gauss are not uncommon. Such a splitting is larger than the ones of concern in this review. However the effect of the dipolar term is very materially reduced by molecular motion in a well-known manner (Bloembergen, Purcell and Pound 1948). In mobile liquids even with a high concentration of magnetic nuclei the actual line width may be only milligauss or less. The reducing effect of random variation in $\mathbf{r}_{ij}$ will not be dealt with but it is obvious that isotropic variation even of θ_{ij} gives $\mathscr{H}_{\text{dsec}} \to 0$ since $\cos^2\theta_{ij} \to \frac{1}{3}$. Indeed we readily show that if the motion has a correlation time τ_c a well narrowed line has width (Bloembergen, Purcell and Pound 1948, Pines and Slichter 1955)

$$\delta H = (\delta H_{\text{rigid}})^2 \gamma \tau_c \qquad \ldots\ldots(9)$$

where δH_{rigid} is the line width in the absence of motion. In terms of frequency this is

$$\delta\nu = (\delta\nu_{\text{rigid}})^2 \tau_c. \qquad \ldots\ldots(9')$$

It is on the whole preferable to work in terms of frequency particularly when comparing different nuclei since γ does not appear in (9′) and similar formulae but we shall use both since the experimentally observed quantity is often a field shift.

For an interaction corresponding to a few gauss in rigid lattice line width and for a mobile liquid at room temperature for which we might have $\tau_c \simeq 10^{-11}$ sec we find $\delta H \sim 10^{-2}$ mG or, for protons, $\delta\nu \simeq \frac{1}{10}$ c/s or T_2 several seconds. A higher viscosity liquid tends to have a longer τ_c and, roughly speaking, liquids with viscosities some one hundred times as high as that of water may have sufficiently rapid molecular motion so as to keep the line width due to dipolar interaction down to about a milligauss. In this situation the line width is usually controlled

by T_1 broadening and precautions such as avoiding paramagnetic impurities must be taken in order to observe the true line width of the material. Since all nuclei other than protons have a smaller magnetic moment (except for tritium) we may expect to detect effects of the order of or even smaller than a milligauss in quite a wide range of liquids. Nuclear spin–rotational moment interactions, important in beams experiments, are also averaged out.

Some of the effects we shall discuss may be observed in solids for the heavier nuclei (Bloembergen and Rowland 1955, Gutowsky and McGarvey 1953) since these tend to have both smaller magnetic moments and larger electronic effects. When motional averaging of dipolar interaction by factors of some 10^4 are involved a closer consideration of equation (7) is required and we note, for instance, that in fact the direct dipolar interaction takes place through the body of the molecule which is a diamagnetic medium which, if it is anisotropic, may leave a residual effect which might well be represented by an isotropic term in the Hamiltonian of the form $-hJ_{dij}\,\mathbf{I}_i.\mathbf{I}_j$.

In the early use of nuclear magnetic resonance as a method for the accurate determination of nuclear moments it was noticed that the resonance frequency for a given nucleus varied according to the material containing it (Knight 1949, Dickinson 1950a). Examples for a large number of nuclei are given by Proctor and Yu (1951). Cobalt (^{59}Co) shows the large variation in resonant frequency of order one per cent as shown below and this is a molecular effect. (More recent work by Freeman *et al.* (1957) confirms the unusually large shifts in cobalt complexes.)

^{59}Co in	$K_3Co(CN)_6$	$K_3Co(C_2O_4)_3$	$Co[C_2H_4(NH_2)_2]_3Cl_3$
Relative resonant frequency	1·0000†	1·0130	1·0073

† Unity by definition.

For protons in water, paraffin, etc., the resonant frequency varies by 1 in 10^6 or more and by some 1 in 10^5 from the value for the free proton (Gutowsky and McClure 1951). Effects are actually much larger for nuclei of higher atomic weight although attention was first attracted to the proton. The differences observed are not explained by small differences in bulk susceptibility of the material. The shift in field may be about

$$\left(\frac{4\pi}{3}-\kappa\right)\chi H_0 \qquad \ldots\ldots(10)$$

for a material of susceptibility χ in a field H_0 and with a shape giving the demagnetization factor κ (Hood, Redlich and Reilly 1954). The determination of the actual magnetic field ' inside ' a material, or at a nucleus, is comparable with the internal field problem in dielectrics and a satisfactory solution is not yet available. In many of the experiments the specimen is a long cylinder perpendicular to the field and so $\kappa = 2\pi$. Since $\chi \simeq 10^{-6}$ for most diamagnetic liquids the correction is small but in a field of 10^4 G may be of order 10 mG. Since diamagnetic liquids of similar type have very similar diamagnetic susceptibilities the effect in comparative experiments would be at least an order of magnitude lower, e.g. 1 mG for protons in organic liquids. In addition to this we shall often be concerned with relative shifts in the same molecule rather than absolute shifts and so bulk corrections will not usually be important. In solution studies it may be a very significant

correction. These macroscopic shifts are nevertheless large enough to make them the basis of a method of measuring macroscopic susceptibilities of liquids (Feher and Knight 1955).

2.2. *Magnetic Shielding in Atoms*

It is readily understood that the field at the nucleus is not the applied field but is usually rather less as a result of the diamagnetic screening effect of the orbital electrons. The field at the nucleus of an atom may be written

$$H_n = H_0 + H_{dia} = H_0(1-\sigma). \qquad \ldots\ldots(11)$$

Owing to the difficulty in obtaining free nuclei of any but the lightest atoms σ cannot be determined experimentally but fortunately it may be calculated with some confidence for atoms. The calculation is similar to that of the diamagnetic susceptibility of an atom. In the magnetic field H_0 the electrons in the atom acquire a superimposed Larmor precession $\nu_L = eH_0/4\pi mc$ and the circular current $i = e\nu_L/c$ e.m.u. of radius a produces a field $2\pi ia^2/r^3$ at the nucleus, i.e.

$$H_{dia} = -\frac{e^2}{2mc^2}\frac{x^2+y^2}{r^3}H_0 \quad \text{per unit charge}$$

whence

$$\sigma_L = \frac{e^2}{2mc^2}\int\frac{x^2+y^2}{r^3}\rho d\tau = \frac{e^2}{3mc^2}\int\frac{\rho}{r}d\tau = \frac{e^2}{3mc^2}V(0) \qquad \ldots\ldots(12)$$

where $V(0)$ is the electrostatic potential at the nucleus which may be obtained from the wave function of the atom by, for example, the Thomas–Fermi method (Lamb 1941) or Hartree's method (Dickinson 1950b).

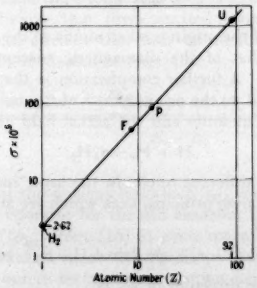


Figure 2. Magnetic shielding plotted against atomic number (using Tables VI, II of Ramsey 1956). The straight line is $\sigma = 2{\cdot}62\times10^{-5}\,Z^{1\cdot35}$.

For the hydrogen atom $\langle 1/r\rangle = 1/a_0$ where $a_0 = \hbar^2/me^2$ is the radius of the first Bohr orbit and so for atomic hydrogen $\sigma = \alpha^2/3$ where α is the fine structure constant $(2\pi e^2/ch = 1/137)$ thus for the hydrogen atom $\sigma = 1{\cdot}78\times10^{-5}$ (for molecular hydrogen σ is actually $2{\cdot}62\times10^{-5}$). For heavier atoms it is larger roughly in proportion to $Z^{1\cdot35}$ (see figure 2).

We note a great difference between σ and χ, the diamagnetic susceptibility, in that χ depends on $\langle r^2 \rangle$ so that it is sensitive to the outer electrons whereas σ depends on $\langle 1/r \rangle$ and hence mainly on the inner electrons. However in comparing σ values the accuracy attainable is such that changes in outer electrons are readily measurable. We note in passing that

$$\sigma/\chi_a = 2\langle r^{-1} \rangle / \langle r^2 \rangle \qquad \ldots\ldots(13)$$

where χ_a is the susceptibility per atom. The diamagnetic shielding field for $H_0 = 10^4$ gauss varies between 262 mG for the proton in H_2 to 116 gauss for uranium. The corresponding frequency shift in hydrogen is 1110 c/s.

2.3. *Magnetic Shielding in Molecules*

2.3.1. *Perturbation methods.*

The theory does not apply to molecules since the spherical symmetry is removed and the field does not induce a simple Larmor precession for an electron under more than one attracting centre. The 'error' for the hydrogen molecule is

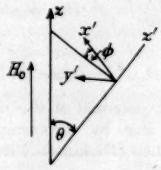


Figure 3.

almost as large as σ itself if the origin is taken at one of the protons. The difference is closely analogous to that of the diamagnetic susceptibility of a molecule as compared with an atom. A further complication is that the field at the nucleus may now evidently depend on the orientation λ of the molecule relative to the field $\mathbf{H}_0$, i.e. $\boldsymbol{\sigma}$ is now a tensor quantity and the actual field at the nucleus is

$$\mathbf{H} = \mathbf{H}_0 - \boldsymbol{\sigma}_\lambda \mathbf{H}_0 \qquad \ldots\ldots(14)$$

and may be different for different nuclei in the same molecule. If the shielding tensor is $\boldsymbol{\sigma}'$ referred to its own principal axes which are at (θ, ϕ) to H_0 (see figure 3) the field at the nucleus is

$$H_0[\sin\theta\cos\phi\,(1-\sigma_x'),\ \sin\theta\sin\phi\,(1-\sigma_y'),\ \cos\theta\,(1-\sigma_z')]$$

and the component in the direction of H_0 is

$$H_0[1-(\sigma_x'\sin^2\theta\cos^2\phi + \sigma_y'\sin^2\theta\sin^2\phi + \sigma_z'\cos^2\theta)]$$

which for isotropic tumbling is

$$H_0[1-\tfrac{1}{3}(\sigma_z' + \sigma_x' + \sigma_y')] \quad \text{i.e. } \sigma = \tfrac{1}{3}(\sigma_z' + \sigma_x' + \sigma_y').$$

For axial symmetry $(\sigma_x' = \sigma_y')$ the transverse field is $H_0 \sin\theta\cos\theta(\sigma_z' - \sigma_x')$ i.e. proportional to the anisotropy of $\boldsymbol{\sigma}'$, a result which is required later. In a nuclear resonance experiment the field at the nucleus is averaged over the motion of the

molecule. This may be a rather restricted rotational motion in a beam experiment but in magnetic resonance experiments on macroscopic material σ is averaged over the isotropic tumbling motion of the molecule which is essential to reduce the direct dipole–dipole interaction and make the line sharp enough to measure σ with sufficient accuracy. Even beam experiments are as yet hardly accurate enough to detect the anisotropy of σ although estimates can be made indirectly (Ramsey 1951). Some estimate of the anisotropy of σ may be obtained from its effect in reducing the spin–lattice relaxation time which is discussed in § 2.3.6.

The calculation of σ_A for a nucleus in a given molecule may be compared with the calculation of the diamagnetic susceptibility of a molecule (Van Vleck 1932). It is found that in order to obtain σ_A one has not only the ordinary Lamb term (integrated of course over all the electrons in the molecule) but also another term which corresponds to the second-order paramagnetism term or ' high frequency ' term of χ. σ_{HF} like χ_{HF} is paramagnetic so that if

$$\sigma = \sigma_L + \sigma_{HF} \qquad \ldots\ldots(15)$$

σ_{HF} reduces the magnetic shielding corresponding to the term σ_L. The term σ_{HF} may be obtained by second-order perturbation theory since the molecular ground state is non-magnetic, and so one requires a knowledge of the excited electronic states of the molecule. The excited states may have electron spin and/or orbital moments. However, the spin moment does not contribute a field at the nucleus since the interaction of the two unpaired spins with the applied field H_0 is equal and opposite, the field H_0 being uniform over the molecule. (A similar argument applies to the calculation of χ_{HF}.) If the field were not uniform there could be a contribution and we shall find that this is the source of another interaction to be discussed in § 3, where the perturbing field is that of the nuclear moment itself. Although it is obvious that these excited states make a paramagnetic contribution to χ, it is less obvious that they do so for σ. Ramsey (1950 a, b, 1952) has shown in a manner described later that

$$\sigma_{HF} = -2\sum_{n\lambda'}{}' \frac{\left\langle 0\lambda\left|\sum_k m_{zk}^0\right|n\lambda'\right\rangle\left\langle n\lambda'\left|\sum_k \frac{m_{zk}^0}{r_k^3}\right|0\lambda\right\rangle + \left\langle 0\lambda\left|\sum_k \frac{m_{zk}^0}{r_k^3}\right|n\lambda'\right\rangle\left\langle n\lambda'\left|\sum_k m_{zk}^0\right|0\lambda\right\rangle}{E_n - E_0} \qquad \ldots\ldots(16)$$

where the origin of r_k is at the nucleus for which σ_{HF} is desired, $m_{zk}{}^0$ is the z axis orbital magnetic moment operator for the kth electron, H_0 being in the Oz direction. The calculation of σ_{HF} from (16) is even more forbidding than it looks because not only are the excited state wave functions usually rather poorly known, but the ones which turn out to be important are not usually the ones which take part in actually observed (e.g. in ultra-violet absorption) transitions.

Since usually $E_n - E_0 \gg kT$, σ_{HF} is temperature independent, but this is not always so (Proctor and Yu 1951). Generally speaking, when χ_{HF} is temperature dependent due to low-lying magnetic excited levels, so is σ_{HF}. Freeman *et al.* (1957) find evidence of a correlation between optical absorption and the shielding in cobalt complexes.

σ_{HF} is zero in an atom because expressions like $\langle 0\lambda|\Sigma_k m_{zk}{}^0|n\lambda'\rangle$ must all be zero by symmetry. Further σ_{HF} tends to reflect the symmetry of the molecule,

e.g. $(\sigma_{HF})_\lambda = 0$ for λ is the axis of a linear molecule since $m_z \sim \partial/\partial\phi$. The expression (16) is simplified if we replace all $1/(E_n - E_0)$ by an average value $1/\Delta E$ and use the closure property (Schiff 1955), and becomes

$$\sigma_{HF} = -(4/3)\langle 0|\Sigma_{jk}\mathbf{m}_j^0.\mathbf{m}_k^0/r_k^3|0\rangle/\Delta E$$

which depends only on ground state wave functions. For hydrogen one finds from experiment ΔE about 20 ev which is comparable with the ionization energy of the hydrogen molecule (16 ev), indicating that highly excited states are involved.

A rather simple explanation of the expression for σ_{HF} is readily obtained. A system with ground state ϕ_0 and one excited state ϕ_n of excitation energy $E_n - E_0$ has a ground state wave function in the magnetic field H_0 along Oz of

$$\phi_0' = \phi_0 + \frac{H_0\langle n|m_z|O\rangle}{E_n - E_0}\phi_n.$$

Thus the magnetic moment of the molecule in H_0 is

$$\langle O'|m_z|O'\rangle = \frac{2H_0}{E_n - E_0}|\langle O|m_z|n\rangle|^2$$

which is known to be positive since χ_{HF} is positive and so the 'internal field' at the nucleus due to this magnetization will be in the direction of H_0. Since an element of magnetic moment $d\mu$ at (r, θ, ϕ) gives a field δH_{HF} at the origin in the Oz direction of $(d\mu/r^3)(3\cos^2\theta + 1)^{1/2} \simeq 2d\mu/r^3$, therefore $H_{HF} \simeq 2\langle O'|m_z/r^3|O'\rangle$,

hence $$\sigma_{HF} = -H_{HF}/H_0 \simeq -4\langle O|m_z|n\rangle\langle n|m_z/r^3|O\rangle/(E_n - E_0), \quad \ldots\ldots(17)$$

which is of the form (16). An order of magnitude for σ_{HF} is obtained if we assume $E_n - E_0$ is the ionization energy $e^2/2a_0$ of the hydrogen atom and

$$\langle O|m_z|n\rangle\langle n|m_z/r^3|O\rangle \simeq \langle O|m_z^2/r^3|O\rangle \simeq \beta^2/a_0^3$$

where β is the Bohr magneton. Thus $\sigma_{HF} \simeq -\alpha^2$ and so for one hydrogen-like electron $\sigma_L \simeq -\sigma_{HF}$. However, in a molecule all electrons contribute to σ_L but only suitably excitable ones contribute to σ_{HF}, and so for heavier nuclei $|\sigma_{HF}| \ll |\sigma_L|$.

Even in the case of the hydrogen molecule calculation of σ_{HF} is very difficult (Hylleraas and Skavlam 1950, Ishiguro and Koide 1954). Estimated values are $\sigma_L = 3{\cdot}21 \times 10^{-5}$, $\sigma_{HF} = -0{\cdot}59 \times 10^{-5}$, so that $\sigma = 2{\cdot}62 \times 10^{-5}$ (these are values averaged over isotropic tumbling). Ramsey (1950 a, b) has obtained a more reliable value for σ_{HF} by using the nuclear spin-rotational moment interaction constant which is known from beam experiments and which is closely related to the expression in (16). This gives $\sigma = 2{\cdot}62 \times 10^{-5}$ (Ramsey 1950 b, Kolsky *et al.* 1952) and almost the same value for D_2, as expected. Experimentally, the difference in shielding of H_2 and D_2 is about 1 in 10^7 (Wimett 1953). In all other molecules σ_{HF} is much less than σ_L. For linear molecules the calculations can be performed with some confidence. The theory has been applied to the relative shielding in HF and F_2 with some success (Saika and Slichter 1954). ^{31}P shielding has been discussed by Muller *et al.* (1956) and Parks (1957).

Often the shielding for a given nucleus is very close to that which would be obtained using the simple Lamb formula referred to the electrons specifically associated with that nucleus although there are some exceptional cases such as the

cobalt complexes. This is because the additional contribution to σ_L in the molecule as compared with the atom tends to be compensated for by the additional term σ_{HF} which is opposite in sign. Thus in the case certainly of heavier nuclei a rather accurate value of σ is readily obtained. However, as we shall see, small *differences* of shielding arising from differences in chemical position within the molecule and from molecule to molecule are readily measured, and for the calculation of these effects a much higher accuracy is necessary.

There are two objections to this method of calculating σ. One is that it requires a detailed knowledge of the excited state wave functions (not necessarily states readily attainable experimentally). Secondly, the method is artificial in that the separation of σ into σ_L and σ_{HF} is an artificial one. This is clearly shown in Ramsey's (1952) treatment. Take the vector potential $\mathbf{A}$ of an electron at $\mathbf{r}$ from a nucleus of moment $\boldsymbol{\mu}$ as

$$\mathbf{A} = \tfrac{1}{2}\mathbf{H} \wedge \mathbf{r} + \boldsymbol{\mu} \wedge \mathbf{r}/r^3 - \tfrac{1}{2}\mathbf{H} \wedge \mathbf{R} \qquad \ldots\ldots(18)$$

where $\mathbf{R}$ is an arbitrary irrotational vector expressing the well-known arbitrariness of the gauge of the vector potential and of course any observable must be independent of $\mathbf{R}$. The first term gives the applied field $\mathbf{H}$ and the second is the field due to the nuclear dipole itself. We then use the Schrödinger equation with a Hamiltonian operator in the well-known form

$$\mathscr{H} = V + \frac{1}{2m}\left(\mathbf{p} + \frac{e}{c}\mathbf{A}\right)^2 \qquad \ldots\ldots(19)$$

where for brevity we do not write the summation over all the electrons and we omit terms involving the electron spin for reasons already discussed. V is the nuclear electrostatic potential. Inserting (18) in (19) and developing it we find

$$\mathscr{H} = V + T + D + L + \mathrm{ON} + \mathrm{HN} + \mathrm{NN} \qquad \ldots\ldots(20)$$

where T is the kinetic and D the diamagnetic (χ_L) energy of the molecule and do not concern us. $L = \beta\mathbf{H}.\mathbf{L}$ where $\mathbf{L}$ is the total orbital moment, i.e. the orbital interaction with the field. $\mathrm{ON} = 2\beta\gamma\mathbf{m}.\mathbf{I}/r^3$ and is evidently a coupling between an electron orbital moment and the nucleus. $\mathrm{HN} = (e^2/2mc^2)(\mathbf{H} \wedge \mathbf{r})(\mathbf{r} \wedge \boldsymbol{\mu})/r^3$ is a term involving the nucleus and the applied field and so is of immediate interest. $\mathrm{NN} = (e^2/2m_e^2)(\boldsymbol{\mu}_i \wedge \mathbf{r}_i/r_i^3)(\boldsymbol{\mu}_j \wedge \mathbf{r}_j/r_j^3)$ is an interaction between nuclear spins which will be of interest in another connection (§ 3.3).

Regarding the last four terms as small perturbations, we see that in first order $L = 0$ since the molecule is diamagnetic and ON is zero for the same reason. HN gives an energy term in first order which is bilinear in μ and H and so is of interest; it may in fact be shown to be σ_L. It is now necessary to go to a second-order perturbation calculation to find another term in μH and the only terms suitable are L and ON and this leads to a term of the form

$$E_{L,\mathrm{ON}} = 2\beta^2\gamma\langle \mathrm{O}|\mathbf{H}.\mathbf{L}|n\rangle\langle n|\mathbf{m}.\mathbf{I}/r^3|\mathrm{O}\rangle/(E_n - E_0) \qquad \ldots\ldots(21)$$

which is indeed the origin of σ_{HF} in equation (16) remembering that this depends on the coefficients of $\mu_i H_j$ in $E_{L\ \mathrm{ON}}$.

Ramsey (1952) has shown that $\boldsymbol{\sigma}$ is independent of choice of $\mathbf{R}$ but that if $\mathbf{R} = 0$ the two terms are precisely the $\boldsymbol{\sigma}_L$ and $\boldsymbol{\sigma}_{HF}$ found previously, so that the

division of $\boldsymbol{\sigma}$ was quite arbitrary. Incidentally this explains why distant electrons do not contribute much to σ even though σ_L depends on $\langle 1/r \rangle$ and so is an extended integral; it is because the long-range contributions to σ_L are compensated by σ_{HF}.

2.3.2. *Variational methods.*

The alternative to a perturbation calculation is to use the variational method and this has been exploited by Hornig and Hirschfelder (1955), Das and Bersohn (1956), McGarvey (1957 a, b) and Stephen (1957 a). The calculations follow the way indicated by Tillieu and Guy's (1956) calculation of diamagnetic susceptibility by a variational method (see also Tillieu 1957).

Using the vector potential (18) we do a variational calculation in which the energy of the molecule in the magnetic field $\mathbf{H}$ is

$$E = E_0 - \boldsymbol{\mu}.\mathbf{H} - \boldsymbol{\mu}.\boldsymbol{\sigma}.\mathbf{H} \qquad \ldots\ldots(22)$$

and is to be minimized by suitable choice of a wave function of the form

$$\phi = \phi_0(1 + \boldsymbol{\mu}.\mathbf{f} + \mathbf{H}.\mathbf{g}). \qquad \ldots\ldots(23)$$

The results of the variational calculation (which is shown also to be gauge invariant (Stephen 1957 a)) are

$$\sigma_{ij} = \langle \mathrm{HN} \rangle - \frac{\hbar^2}{m}\left[\left\langle 0 \middle| \sum_k \nabla_k f_i \nabla_k g_j \middle| 0 \right\rangle + \left\langle 0 \middle| \sum_k \frac{2mk_i}{r_k^3} g_j + m_{kj} f_i \middle| 0 \right\rangle\right] \qquad \ldots\ldots(24)$$

where σ_{ij} are the components of the shielding tensor and HN is as defined in § 2.3.1, together with the variational equations for $\mathbf{f}$ and $\mathbf{g}$. We therefore again get two terms (corresponding to σ_L and σ_{HF}). For a tumbling molecule only the isotropic $\sigma = \frac{1}{3}(\sigma_x' + \sigma_y' + \sigma_z')$ is required (see § 2.3.1).

The general computational difficulty remains that the field for each nucleus in the molecule must be found independently since the matrix elements refer to that nucleus as centre and are not readily transferred to another centre although we shall mention some useful approximations.

We now require the ground state wave function ϕ_0 and the usual approximate ones of molecular theory may be used and advantage taken of wave functions developed for the calculation of other molecular properties, although these may not be of satisfactory accuracy in the respects required in this calculation. The use of the conventional approximate wave functions, molecular orbitals ($\phi_0 = \prod_{k=1}^{2n} \phi_{0k}$ where ϕ_{0k} is a one-electron wave function) or valence bond ($\phi_0 = \prod_{A,B=1}^{n} \phi_{AB}^{(1,2)}$ where $\phi_{AB}^{(1,2)} = \phi_A^{(1)}\phi_B^{(2)} + \phi_B^{(1)}\phi_A^{(2)}$) leads to considerable simplification in that the relations break down to sums of one-electron type expressions. More refined wave functions can be used for better known molecules like H_2. A valuable simplification is that for an electron not localized on the nucleus in question the $1/r^3$ appearing in the expressions (e.g. equation (24)) may be replaced by an average

value d of the centre of gravity of the electron cloud from the nucleus, and the contribution to $\boldsymbol{\sigma}$ is

$$\boldsymbol{\sigma} = d^{-5}[d^2\boldsymbol{\chi} - 3\boldsymbol{\chi}.\mathbf{dd}]$$

where $\boldsymbol{\chi}_{ij}$ is known for a number of cases (Tillieu and Guy 1956) (cf. equation (13)).

Early use of the symmetry properties of the molecule leads to considerable simplification. For a linear molecule with its axis in the Oz direction $\Sigma_k m_{kz}\phi_0 = 0$ since $m_z \sim \partial/\partial\phi$ and ϕ_0 must be independent of ϕ; this gives $f_z = g_z = 0$, i.e. no 'excitation' of the molecule by $\mathbf{H}$ or $\boldsymbol{\mu}$ and hence no contribution to σ_{HF} or more precisely $\sigma_{zzHF} = 0$. Similarly, symmetry considerations severely restrict the form of $\mathbf{f}$ and $\mathbf{g}$. Stephen (1957 a) and McGarvey (1957 b) have made a number of calculations for the hydrogen molecule with the results shown below. (N.B. The results are given for σ_L and σ_{HF} calculated at the proton, although the calculations by Stephen were mostly made with respect to the centre of gravity of the molecule.)

The Hydrogen Molecule

Wave function	$\sigma_L \times 10^5$	$-\sigma_{HF} \times 10^5$	$\sigma \times 10^5$	Author
Valence bond	—	0·53	2·69	Stephen
Molecular orbital	—	0·50	2·64	(1957a)
Weinbaum (1933)	—	—	2·91	
Valence bond	2·78	0·15	2·63	
Molecular orbital	2·83	0·13	2·70	McGarvey
Wang (1928)	2·91	0·17	2·74	(1957b)
Molecular orbital	2·44	0·78	3·22	Hameka (1958)
Ramsey (1950 a, b, 1951, 1952) (experiment and theory)	3·24	0·56	2·68	

Absolute *experimental* values of σ, as opposed to shifts, are usually based on Ramsey's calculated value for the hydrogen molecule using spin–rotation data.

The calculations of σ_L and σ_{HF} by Stephen and McGarvey differ because they used different variational functions. However, the resulting value of σ is less in error than σ_L and σ_{HF} and so there is less concern over the application of the method to more complicated molecules where the wave functions are less well known.

In larger molecules the questions of covalent and ionic bonding arise in the usual formalism. Results for several hydrides (McGarvey 1957 b) are given below and the experimental values in figure 4.

	HI	HBr	HCl	HF	H_2Se	H_2S	H_2O
$\sigma \times 10^5$ calculated (gas)	4·00	3·66	3·05	0·99	3·94	3·59	1·77
Experiment (liquid) (Gutowsky and Hoffman 1951)	4·23	3·44	3·11	2·31	3·38	3·06	2·7

The disagreement for HF and H_2O does not appear to be explained by hydrogen bonding in the liquid, otherwise the theoretical results are satisfactory. It may be that chemical shift is as good a means as any of measuring ionic character!

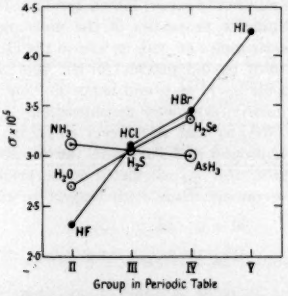


Figure 4. Proton magnetic shielding in simple hydrides showing effect of change in size of heavy atom (Gutowsky and Hoffman 1951).

The proton shielding in hydrocarbons has been calculated by Stephen (1957 a) and requires as a preliminary the value for the C–H bond proton for various types of carbon bonding.

$\sigma \times 10^5$	$-\overset{\vert}{\underset{\vert}{C}}-H$	$=\overset{\vert}{\underset{\vert}{C}}-H$	$\equiv C-H$
Calculated C–H only	2·67	2·65	2·67
Calculated for molecule	Methane 2·65	Ethylene 2·61	Acetylene 2·74
Experiment (liquid) (Meyer *et al.* 1953)	2·84	2·57	2·86
Shift relative to methane; theory (Stephen)	0	−0·04	+0·09
Experiment (gas) (Schneider *et al.* 1958)	0	−0·52	−0·14

The figures in line 3 of the table incorporate a correction to the published ones (Stephen, private communication).

The table shows that the main contribution to σ comes from the electrons in the bond at the nucleus concerned for these molecules. Elementary considerations of electron distribution in the series methane, ethylene, acetylene would suggest that the shielding should be in the same order, but it is not. This question is taken up again below.

It has not proved possible to deduce any simple rules for the importance of various parameters in chemical shielding from the theories particularly for protons because of the diverse effects of most parameters (see McGarvey's conclusions 1957b).

The experimental results, however, show empirical correlation with a number of other properties such as bond distance, quadrupole coupling (Meyer and Gutowsky 1953), ring size in cyclic compounds (Gutowsky *et al.* 1954), Hammett's substituent sigma value (Gutowsky *et al.* 1951, 1952), ionic character, electronegativity (Shoolery 1954), size of attached atom and covalency (Gutowsky and Hoffman 1951) (see figure 4) etc. The empirical approach has been most profitable for ^{19}F shifts (Gutowsky 1956).

Calculations of proton shielding are less simple than for heavier nuclei since the effect of distant electrons is relatively much more important. This fact no doubt explains a number of anomalous proton shielding values (McConnell 1957 a).

2.3.3. *Circulating currents.*

Another approach which has the virtue of giving a more physical picture of the shielding values, and will therefore facilitate the making of simplifying assumptions, has been exploited by Pople and others (Pople 1956, 1957, Bernstein *et al.* 1956). The diamagnetic shielding in atoms may be regarded as due to the Larmor circulation of electrons induced by the applied field. This idea can be generalized to a molecule where one might distinguish three types of induced circulating current :

(*a*) a local diamagnetic circulation about each nucleus,

(*b*) a paramagnetic circulation which can be split into circulations about each nucleus, and

(*c*) interatomic currents flowing in molecular paths.

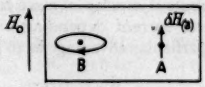


Figure 5.

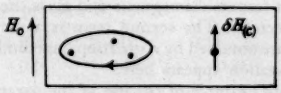


Figure 6.

(*a*) gives a shielding not much different from σ_L for the atom and is the principal contribution to the shielding except for light nuclei. This suggests that values should increase with electron density near the nucleus and hence should correlate with ionic character or electronegativity, but the facts are not quite so simple (Meyer *et al.* 1953), presumably because of simultaneous changes in (*b*) and (*c*). We note that a diamagnetic circulation about another nucleus B may cause a paramagnetic field $\delta H_{(a)}$ at the nucleus A under study (see figure 5, which shows conventional current). Although the effect may be small it may be important in comparative studies. However, with molecular tumbling the average effect is non-zero only as a result of diamagnetic anisotropy and so is further reduced.

(*b*) arises from the mixing of the ground and excited states to give the wave function in the applied magnetic field. This may be regarded as giving rise to further currents which give a paramagnetic field at the nucleus and may therefore be called paramagnetic currents. These currents circulate in the molecule as a whole but may with some loss of rigour be split up into currents circulating about each nucleus in the molecule. This leads to some simplification because the

paramagnetic current associated with a nucleus other than the one in question will produce a field at the test nucleus which can be calculated approximately by replacing that local circulating current by a point dipole. The field due to paramagnetic circulating currents about another atom is again only effective if anisotropic because of molecular tumbling. In general, paramagnetic circulations are more likely to be anisotropic than diamagnetic ones. For example, there will be no paramagnetic circulation about the axis of a linear molecule. For (*c*) we note that there may be interatomic currents flowing round a molecular path. Such currents, although essentially diamagnetic as far as susceptibility is concerned, may well be paramagnetic in shielding effect since, for instance, the current may produce a field in the direction of H_0 at a nucleus, particularly if it is outside the current (see figure 6). The molecular currents are obviously going to be most important for such cases as benzene rings where a path of large dimensions is possible, and we shall deal with these separately below. The formal analysis uses the Hamiltonian (19) together with the electronic current operator

$$\mathbf{j}(\mathbf{r}) = -\frac{e}{m}\sum_j \left[\mathbf{p}_i + \frac{e}{c}\mathbf{A}(\mathbf{r})\right]\delta(\mathbf{r}-\mathbf{r}_j)_{\text{sym}} \qquad \ldots\ldots(25)$$

which is the operator generalization of the expression $\mathbf{j} = -(e/m)\mathbf{p}$ to the case with magnetic fields.

With a wave function $\phi = \phi_0 + i\Sigma_k c_k \phi_k$ the expectation value of the current is

$$\langle \mathbf{j} \rangle = \langle 0|\mathbf{j}|0\rangle + i\sum_k \{c_k\langle 0|j|k\rangle - c_k^*\langle k|j|0\rangle\}$$

the first term is diamagnetic and gives the (*a*) currents if overlap of wave functions is neglected. The second term is a paramagnetic current depending on the c_k which are obtained by a variational method. No particular advantage as to physical interpretation appears here.

As an example of the use of the paramagnetic current, Pople (1957) estimated the paramagnetic moment on each carbon atom in acetylene (CH≡CH), since the orbitals and the excited states of this molecule can be predicted with some confidence. Symmetry considerations restrict considerably the excited states which have to be considered. The moment on the carbon atom of acetylene is found to be $\mu_c = 3e^2\hbar^2 H_0/4m^2c^2\Delta\bar{E}$, where $\Delta\bar{E}$ is a mean excitation energy and when the molecular axis is at right angles to H_0. Since we have a linear molecule $\mu_{x'} = 0$ and $\mu_{x'} = \mu_{y'} = \mu_c$. Owing to the molecular tumbling the actual Oz field at the near proton is $\frac{1}{3}(2\mu_{cz} - \mu_{cx} - \mu_{cy})/R^3$ which in this case is simply $-\frac{2}{3}\mu_c/R^3$, R being the C–H distance, as we readily see from figure 7 on performing the usual averaging. The result is a diamagnetic contribution to σ of $1{\cdot}0\times10^{-5}$ which is encouraging because ordinary electron density consideration in the series ethane (CH_3-CH_3), ethylene ($CH_2=CH_2$), acetylene (CH≡CH) suggest that σ in acetylene is in fact about $0{\cdot}76\times10^{-5}$ too large.

Again the proton shielding in the bonds N–H, O–H and F–H is about the same (in the vapour) as in $-CH_3$, whereas the electron concentration near the proton must be very different. The effect is thought to be due to paramagnetic circulations on the neighbour atoms and preliminary calculations tend to confirm this (Pople 1957).

On the other hand, there are cases where a considerable change in shielding occurs without a great change in electron density, for example the differences in shielding as between a paraffinic methyl proton ($-C-CH_3$) and an aldehydic proton (C−CHO) is $0{\cdot}89\times10^{-5}$, which is a considerable proportion of the diamagnetic shielding of the hydrogen atom ($1{\cdot}8\times10^{-5}$) and so can hardly be due to changes in electron distribution alone.

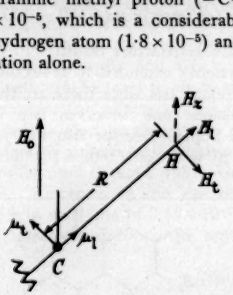


Figure 7. $\mu_1 = \mu_{cz'}\cos\theta$, $\mu_t = \mu_{cz'}\sin\theta\cos^2\phi + \mu_{cy'}\sin\theta\sin^2\phi$,
$H_1 = 2\mu_1/R^3$, $H_t = \mu_t/R^3$, $\bar{H}_z = \frac{1}{2}(2\mu_{cz} - \mu_{cz} - \mu_{cy})/R^3$.
θ is the angle between H_1 and H_z.

Many shifts are very small compared with the total σ value and so only calculations of the highest accuracy (say to 1%) give any interpretation of the experimental results. For example, the shifts relative to liquid water at room temperature for variously bonded methyl protons, X, are :

Proton X	C–CH_3	N–CH_3	O–CH_3	F–CH_3	CH_2 cyclic	C_6H_6
$(\sigma_X - \sigma_{H_2O})\times10^5$	+0·41	+0·25	+0·16	+0·23	+0·35	−0·19

We note that a typical difference, $\Delta\sigma\times10^5 = 0{\cdot}1$, at $H_0 = 10^4$ gauss is 10 mG, whereas the actual shifts with respect to a bare proton are some 300 mG ($\sigma_{H_2O} = +2{\cdot}62\times10^{-5}$, Thomas 1950).

In benzene the shielding is smaller than in the 'normal' cases. Such an effect in benzene and the aromatic hydrocarbons is readily explained by the well-known mobility of the π electrons and is a case where the contribution (c) from molecular currents is readily calculated in a manner similar to Pauling's (1936) celebrated calculation of the diamagnetic anisotropy of the aromatic molecules.

The contribution of the π electron molecular currents is illustrated by the difference between ethylenic and aromatic proton shielding (Pople 1956): $\sigma(\text{benzene}) - \sigma(\text{ethylene}) = -0{\cdot}14\times10^{-5}$, whereas one would expect the shifts to be almost equal. In the applied field H_0 perpendicular to the plane of the benzene the π electrons will have an imposed circulation at the Larmor frequency $\nu = eH_0/4\pi mc$ and the current for the six π electrons is $i = 6ev/c = 3e^2H_0/2\pi mc^2$. If we assume the current is circular of radius a (the C−C distance), which is that of the hexagon, this corresponds to a magnetic dipole at the centre, $\mu = \pi a^2 i = 3e^2H_0\,a^2/2mc^2$ opposing the applied field, and produces a paramagnetic field at the protons of $\mu/(a+b)^3$ where b is the C−H distance. When the molecular axis makes an angle θ with H_0 the current and the component of the local

field in the direction of H_0 are both reduced by a factor $\cos\theta$, and so for a tumbling molecule we get a reduction by $\langle\cos^2\theta\rangle = 1/3$, giving

$$\Delta\sigma = (e^2/2mc^2)\,[a^2/(a+b)^3] = 0{\cdot}17\times10^{-5}$$

for $a = 1{\cdot}4$ Å and $b = 1{\cdot}1$ Å, a satisfactory result.

The same method is readily extended to molecules with more than one ring which constitutes a more severe test since there are then several types of protons. For instance, in naphthalene one observes two lines with equal intensity (Bernstein, Schneider and Pople 1956, see also 1957) presumably due to protons of types 1 and 2 (figure 8) with $\Delta\sigma = 0{\cdot}02\times10^{-5}$ (1956) or $0{\cdot}033\times10^{-5}$ (1957).

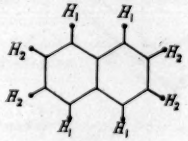


Figure 8.

The calculated π electron shifts for the two positions, assuming two equal circulatory currents, are $\Delta\sigma_2 = 0{\cdot}164$ and $\Delta\sigma_1 = 0{\cdot}201$ or $\Delta\sigma = 0{\cdot}037\times10^{-5}$, which is in excellent agreement with experiment and makes it reasonable to assign the line with the least shielding to protons of type 1. The anti-shielding is greater for protons 1 because they are closer to the current in the other hexagon. (If one nucleus is less shielded than another the resonance will occur at a lower external field for a given frequency.) The method is also successful for a series of multiple ring aromatics with the simple assumption that all ring currents are the same (as that of benzene). The method can also be used to calculate the relative shifts as between protons in different compounds after due correction has been made for changes in macroscopic susceptibility, although this is not as successful, larger hydrocarbons tending to have larger shifts than expected. A refinement of the calculation (Hazato 1957) gives rather better agreement. McConnell (1957 a), in an illuminating article, has shown how this theory of the long-range shielding contributions is related to the second-order perturbation theory described earlier.

2.3.4. *Intermolecular effects.*

We have so far compared the calculated free molecule shielding with that observed in the liquid, assuming them to be the same except for the bulk diamagnetic correction. However, owing to the long-range effects described above, among others, there will be intermolecular effects which will depend on magnetic and shape anisotropy. Aromatic ('flat') molecules should show an appreciable effect of this kind as found by Corio and Dailey (1956 a, b). Solvent effects must be well known when an 'internal' reference substance is used to avoid necessity of two sample coils or samples and to ensure the reference and sample being in the same magnetic field. The reference is more often in a tube coaxial with the sample tube.

Generally the chemical shift is indeed a molecular property as shown by solution experiments (Corio and Dailey 1956 b). For example, the para proton shielding in monosubstituted benzenes is almost the same in the pure liquid as in solution in a number of solvents. However, this is not always true. Effects appear clearly in work on solutions as solvent-dependent shifts which are not explained by bulk shielding (Bothner-By and Glick 1956, 1957). These discrepancies, although very small, can be used to study the structure of the liquid.

Exceptionally large intermolecular effects occur due to hydrogen bonding which evidently both changes the electron density on an already rather exposed proton and puts it in a highly 'anisotropic' situation. Most work has been done on solutions (e.g. Arnold and Packard 1951, Gutowsky and Saika 1953, Huggins *et al.* 1955, 1956). Ammonia in the gas and the liquid phase (Ogg 1954) shows a large difference in shielding of $0{\cdot}15 \times 10^{-5}$ with the higher shielding in the gas. Such observations must of course be reproduced by any theory of

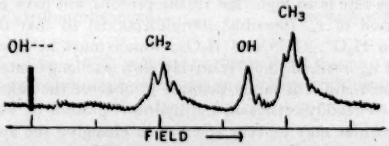


Figure 9. Proton signal for ethanol (CH_3CH_2OH) in carbon tetrachloride. The heavy line indicates the position of the hydroxyl proton signal in pure ethanol (Cohen and Reid 1956).

the hydrogen bond. We discuss ammonia further in § 3, but note in passing that these measurements tend to exclude electron exchange in the hydrogen bond, suggesting that it is electrostatic in nature. Only few measurements have been made on gases because of the relatively weak signals.

Cohen and Reid (1956) have studied the shielding of 'exposed' protons. The dilution shift of hydroxyl protons may be as high as 200 c/s at 40 Mc/s ($\Delta\sigma = 0{\cdot}5 \times 10^{-5}$). They find a striking effect in the classic example, ethyl alcohol. The hydroxyl proton line moves from a field lower than both the methylene and methyl protons to a position above the methyl protons on dilution to 0·05 molar in carbon tetrachloride (figure 9) where it is known that dissociation to monomer is almost complete (Mecke 1950, see also Becker *et al.* 1958). Rise in temperature and dilution have similar effects (see also Packard and Arnold 1951, Arnold and Packard 1951, Liddel and Ramsey 1951). On the other hand, the aldehydic $\left(-C\begin{smallmatrix}\diagup H\\ \diagdown O\end{smallmatrix}\right)$ proton line does not move, suggesting that the association which is known to exist is not by hydrogen bonding but is probably a dipole–dipole coupling associated with the carboxy ($-C{=}O$) group (Schneider and Bernstein 1956). A large difference in proton shielding as between liquid and gaseous water shows the importance of hydrogen bonding in this material (Ogg 1957). See also Diehl 1957. For differences between signals in gas and liquid, see also Schneider *et al.* (1958). We can conclude that in most cases the chemical shift in a liquid is a molecular property, and in cases where it is not one may usually be forewarned

of an anomaly by results from other fields, as in the case of hydrogen bonding. Intermolecular effects are dealt with more generally by Stephen (1958).

The effect of chemical exchange of nuclei on the line structure may be quite marked. If the nucleus has possible sites in two different molecules, e.g. $AH^* + BH \rightleftharpoons AH + BH^*$, then generally the shielding will be different and two lines will be observed with shift $\Delta\sigma$ say. However, if the nucleus exchanges position rapidly enough it will see an average field and a single line at a position depending on the time spent in each site. The changeover occurs when the exchange frequency has the value $\nu_{ec} \simeq \gamma\Delta\sigma H_0/2\pi$. The single line observed in water in which we have the possibilities, among others, of

$$H_2O + H_2H^*O^+ \rightleftharpoons H^*HO + H_3O^+ \quad \text{and} \quad OH^- + H^*H_2O^+ \rightleftharpoons H^*HO + H_2O$$

is due to the rapid exchange of protons between the ionic and molecular forms. Although the actual concentration of these ions in pure water is very low ($\simeq 1$ in 10^7), the exchange rate is so high that all the protons will have exchanged several times in the period of ν_{ec} (possibly 10 milliseconds) so that the differences in shielding between H_3O^+, OH^- and H_3O^-, which must be appreciable, are not observed. Since ν_{ec} is small, only relatively slow exchange rates on a molecular scale are measurable and it is rarely possible to observe the changeover situation. Since exchanges are readily catalysed, e.g. hydroxyl protons by water, even minute quantities of impurities may be very effective in changing the apparent shielding. The intermediate condition line shapes are more difficult to calculate (Gutowsky and Saika 1953, Gutowsky, McCall and Slichter 1953, Grunwald, Loewenstein and Meiboom 1956, Solomon and Bloembergen 1956, Kaplan 1958). An approximate calculation of the exchange modified line shapes may be made by modifying the Bloch equations. Consider two situations in which the Larmor frequency differs by $\Delta\nu = \gamma H_0\Delta\sigma/2\pi$ and allow random changeover from one situation to the other with a mean lifetime in one state or the other of $\tau\ (= 1/2\pi\nu_e)$. The line shape (represented by the transverse rotating magnetization $u + iv$) is given by (Gutowsky, McCall and Slichter 1953)

$$u + iv \propto \frac{-\tau[2 + (1/T_2 - 2\pi i D\nu)\tau]}{(1 + \alpha_+\tau)(1 + \alpha_-\tau) - 1} \qquad \ldots\ldots(26)$$

where $\alpha_\pm = 1/T_2 - 2\pi i(D\nu \pm \frac{1}{2}\Delta\nu)$ and $D\nu$ is the distance from resonance in the mean field $H_0(1 - \frac{1}{2}\Delta\sigma)$. Equation (26) gives two separate lines with widths $1/T_2$ separated by $\Delta\nu$ when $\tau \gg 1/2\pi\Delta\nu$ changing over and broad lines when $\tau \simeq 1/2\pi\Delta\nu$ and one line of width $1/T_2$ when $\tau \ll 1/2\pi\Delta\nu$.

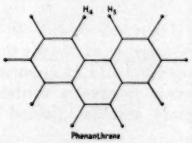
Even when $\nu_e \ll \nu_{ec}$ and the line shape is unaffected the existence of exchange can still be inferred because it forms a link between two apparently independent signals (McConnell and Thompson 1957). If, for instance, one saturates one of the lines the other will saturate because the populations of the two systems of levels must tend to be the same provided the exchange is rapid enough for the effect not to be nullified by spin–lattice relaxation of the populations. Further, saturation of one line tends to saturate the other at the exchange rate, which can therefore be measured in this or some equivalent manner. This method has been used for investigating the behaviour of the ammonium ion (NH_4^+)

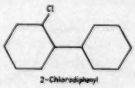
signal, where however the line splitting was due to another cause, but it might be applicable to the measurement of slow exchange rates.

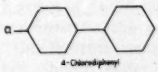
The determination of rate processes by high resolution magnetic resonance is discussed by Phillips (1958).

2.3.5. *Intramolecular effects.*

There is evidence of an effect on the magnetic shielding in some special cases where the nuclei are unusually close. There may be both an electrostatic effect and a purely steric effect on the electron distribution (wave functions) which can lead to abnormal chemical shifts. Such effects have been found, for instance, in O–diCF$_3$ trifluorobenzene for the CF$_3$ ^{19}F signal relative to the 'unstrained' CF$_3$ trifluorobenzene (Gutowsky *et al.* 1952). Again, Bernstein and Schneider (1956) and Reid (1956) found that the protons on the 4 and 5 positions of phenanthrene have an unusually large shift possibly due to steric repulsion of these protons, or perhaps due to ring currents (Pople, private







communication). Similar effects are found in other multi-ring aromatics (see also Reid 1957). Tiers (1956) found a paramagnetic shift in the ^{19}F signal in compounds of the form XCF$_2$Y which is said to be due to a steric effect. An approximate theory (Frank and Gutowsky 1958 a) investigates the effect on the magnetic shielding of a nucleus of an adjacent electric charge via the effect of the electric field on the wave functions which therefore modifies both σ_L and σ_{HF}. For protons a unit charge at 1·5 Å produces a paramagnetic effect of $\Delta\sigma \simeq 0{\cdot}22 \times 10^{-5}$, which is appreciable. For O–diCF$_3$ trifluorobenzene one gets $0{\cdot}22 \times 10^{-5}$, compared with the experimental $0{\cdot}33 \times 10^{-5}$. It would appear that under conditions of steric hindrance the electrostatic field may appreciably affect magnetic shielding.

Marshall and Pople (1958) have calculated the effect of an electric field on the shielding in a hydrogen atom, the results of which are more generally relevant.

Steric effects within molecules may change not only the electron distribution but other parameters such as the internal mobility and even distort the molecule so that certain nuclei are no longer equivalent. Thus 4-chlorodiphenyl shows a singlet proton line, whereas 2-chlorodiphenyl has a double line as a result of 'twist' of the molecule (Brownstein 1958). This should increase the number of

lines observed as well as shifting them from their expected positions. Other steric effects have been observed by Kumler *et al.* (1958).

On the other hand, internal motion of the molecule, or internal rotation tends to reduce the number of lines in the spectrum as well as possibly modifying the position of lines. If the shielding differs by $\Delta\sigma$ in two positions and the change-over from one position to another is more rapid than $\gamma\Delta\sigma H_0/2\pi$ per second on average, the nucleus sees an average shielding $(1-\alpha)\Delta\sigma$ where $0 \leqslant \alpha \leqslant 1$ and α is the proportion of time spent in the second position. In favourable cases the

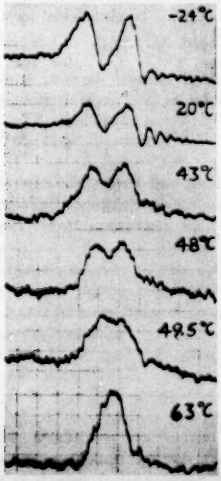


Figure 10. Adjacent methyl proton spectrum in N-N-dimethyl acetamide ($CH_3CON(CH_3)_2$) showing effect of temperature. Total sweep is 19·5 c/s at 17·73 Mc/s (Gutowsky and Holm 1956).

mean lifetime of a state can be found (Bottini and Roberts 1956) and even a value for the activation energy of the process. The effect of internal rotation has been investigated by Gutowsky and Holm (1956), Drysdale and Phillips (1957) and Shoolery and Crawford (1957). The method is, generally speaking, sensitive to only rather slow rates of internal rotation since the shifts are small and these occur in relatively complex molecules. The internal rotation in simple molecules would be of more fundamental interest. Figure 10 shows a steric effect in N-N-dimethyl acetamide (Gutowsky and Holm 1956) where the two adjacent methyl proton signals are slightly displaced at low temperatures, suggesting that the most stable form of the molecule is planar when the methyl groups are different because they are cis and trans to the oxygen. Raising the temperature converts the doublet into a singlet as the internal conversion rate rises through

the critical value ($\simeq 4$ c/s). These effects are usually complicated by the indirect coupling discussed in § 3. The cis–trans isomerism of the alkyl nitrites (RCH_2ONO) has been investigated in a similar way (Phillips *et al.* 1957).

2.3.6. *Anisotropy relaxation.*

A possible way of measuring the anisotropy of the magnetic shielding is from its contribution to the spin–lattice relaxation in liquids (McConnell and Holm 1956, Gutowsky and Woessner 1956). When the shielding is not isotropic the field at the nucleus varies in magnitude and direction with the thermal reorientation of the molecule in the liquid. The nucleus is therefore subjected to a transverse time-varying magnetic field. Since this field has a substantially random variation it will have a Fourier component at the appropriate frequencies (ν_r and $2\nu_r$) to produce spin–lattice type spin flips. It therefore tends to reduce T_1 below its normal value, which is usually due to fluctuating fields generated by moving magnetic nuclei in the same or neighbouring molecules. The magnitude of the transverse shielding field is $\cos\theta \sin\theta\,(\sigma_{z'}-\sigma_{x'})H_0$ (see § 2.3.1) for an axially symmetric shielding tensor σ referred to its principal axes in the molecule and by analogy with the Bloembergen, Purcell and Pound theory (1948) one readily shows that the contribution to the spin–lattice relaxation is given by

$$\left(\frac{1}{T_1}\right)_\sigma = \frac{2}{15}\gamma^2 H_0^2 (\sigma_{z'}-\sigma_{x'})^2 \frac{\tau_c}{1+4\pi^2\nu_r^2\tau_c^2}.$$

Since in liquids $2\pi\nu_r\tau_c \ll 1$, $(1/T_1)_\sigma$ is proportional to H_0^2, whereas other contributions should be independent of H_0. It is found that in a series of fluoro-chloromethanes (CH_3FCl, $CHFCl_2$, CHF_2Cl) and in 1,3,5-trifluorobenzene the proton spin–lattice relaxation times are noticeably longer than the fluorine ones (see table below), whereas from dipolar interaction alone they should differ by a

	T_1 (sec) for ^{1}H		T_1 (sec) for ^{19}F		R_{20}	R_{calc}
	20 Mc/s	26·5 Mc/s	20 Mc/s	26·5 Mc/s		
CH_3FCl	34·6	34·4	9·11	8·69	3·80	0·44
$C_6H_3F_3$	26·0	26·7	16·7	15·6	1·56	0·93

$R_{20} = (T_1$ for ^{1}H$)/(T_1$ for ^{19}F$)$ at 20 Mc/s; $R_{calc} = R$ calculated for dipolar interaction.

factor of order unity. The proton relaxation times appear to have roughly the dipole interaction value. The fluorine shielding is much larger than the proton shielding but it is not self-evident that the anisotropy is much greater. The anisotropy probably depends rather on the order of magnitude of the chemical shifts since these are more closely connected with the outer or valence electrons which are less likely to be in closed shells than the inner ones and so may be expected to make the principal contribution to the anisotropy of the shielding. The variation of T_1 with the applied field shows the correct trend. Admitting a number of approximations, it has been calculated (Gutowsky and Woessner 1956) that in 1,3,5-trifluorobenzene $\sigma_{z'}-\sigma_{x'} \simeq 7{\cdot}5\times10^{-4}$. Using the ^{19}F measurements on trifluorobenzene only, one finds $(\sigma_{z'}-\sigma_{x'})^2\tau_c \simeq 2{\cdot}5\times10^{-28}$. τ_c is not known but dielectric measurements (Curtis *et al.* 1952) on the not vastly dissimilar

chlorobenzene give a dielectric relaxation time $\tau_D \simeq 10^{-11}$ second, and assuming $\tau_c = \tau_D/3$, we find $\sigma_{zz} - \sigma_{xx} \simeq 8 \times 10^{-4}$. An anisotropy as large as 10^{-4} could arise from the second-order paramagnetism associated with a covalent p electron on the fluorine atom (Ramsey 1950 a, b, 1952). This is obviously a rather indirect measure of shielding anisotropy and the interpretation needs care. It is just possible that some of the relaxation in the halogen substituted methanes could originate in anisotropy of the indirect coupling which is discussed in § 3. Effects due to the anisotropy of indirect coupling are not otherwise observed in liquids but are of some importance elsewhere (Solomon 1955, Bloch 1956, Solomon and Bloembergen 1956).

2.3.7. *Chemical structure and analysis.*

Since the nuclear signal depends on the position and number of nuclei in the molecule it can be used to determine molecular structure in favourable cases. Usually it is rather a question of deciding between several alternatives in which it may be possible to put the problem in a form such as: Is there a methyl group present or not?

An illustrative example is that of diketene (Bader *et al.* 1956) for which the possible molecular structures are

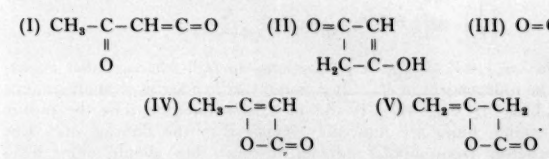


Figure 11 shows the proton signal. Apart from position, the forms I to V should show integrated line intensities, 3,1; 2, 1,1; 4; 3,1 and 2,2, respectively. The spectrum is nearest to 2,2 and further the positions are the ones expected for methylene and vinyl protons and no signals are found for protons of the types

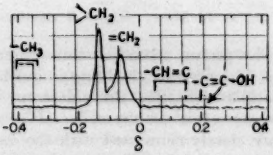


Figure 11. Proton spectrum of diketene showing the presence of only methylene (—CH_2) and vinyl (CH_2=) protons. The expected (see figure 12) positions of other relevant chemical types of protons is indicated (Bader *et al.* 1956).

required for the structures other than V. We may note how much simpler is this than, say, an infra-red analysis, although it must be admitted that this is an unusually favourable case. It is comparatively rare for the spectrum to depend only on the chemical shifts and it is more usually complicated by additional

structure discussed in § 3 which may or may not be helpful. Other examples are given by Corey *et al.* (1955), Gripenberg (1956), Lemieux *et al.* (1957, 1958) and Nachod (1958).

Since the nuclear resonance signal depends on the position of the nucleus in the molecule it can evidently form the basis of a method of chemical analysis and a great deal of the interest in high resolution nuclear resonance stems from this fact. No very useful analytical purpose is served in collecting numerical shielding data alone because most experimental results are complicated by many effects and in particular an important one discussed in § 3. The shielding, as we have seen, may depend on a number of physical parameters and the observed signal may even depend radically on the magnetic field (see figure 20).

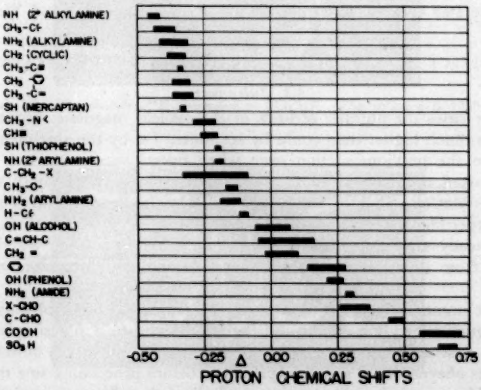


Figure 12. Proton resonance positions relative to water for a number of environments in organic molecules which could be used for analytical purposes (Meyer *et al.* 1953).

High resolution magnetic resonance spectra can, however, when all these possibilities are taken into account, serve as a method of identification of chemical species, particularly for organic molecules. It can also be used in quantitative work (Williams 1958) since the relative—or even absolute—intensities of the signal from nuclei in a given position may serve to indicate the quantity of a given compound present. The method is rather far from the present-day automation of analysis by infra-red absorption. A catalogue of proton spectra of hydrocarbons at 30 Mc/s has been assembled (Williams and Chamberlain of Humble Oil Co., unpublished). Chamberlain (1959) has discussed in great detail the use of high resolution proton spectra for the determination of the structure of organic molecules. See also Gutowsky (1956).

Tables of shielding values are given in the following references but the remarks above should be borne in mind. A typical dictionary of line positions for analysis purposes is shown in figure 12 where experimentally it is convenient to plot

$\Delta = (H_R - H_X)/H_R \times 10^5$ where H_R is the field at resonance for a reference substance (water, say, for protons) and H_X is the field for resonance of the nuclei of type X. The precision of such charts may be very considerably increased by specifying the chemical nature of the proton position more accurately (see Chamberlain 1959).

Shielding data for the following nuclei will be found in the references given: ^{1}H (Gutowsky and Hoffman 1951, 1952), ^{19}F (Gutowsky and Hoffman 1950, Gutowsky *et al.* 1951, 1952, 1953), ^{31}P (Gutowsky *et al.* 1953, Gutowsky and McCall 1954, Van Wazer *et al.* 1956, Muller *et al.* 1956), ^{13}C (Lauterbur 1957, Holm 1957), ^{14}N (Holder and Klein 1955, Masuda and Kanda 1953, Ogg and Ray 1957 a, Schmidt *et al.* 1958), ^{17}O (Weaver *et al.* 1955), ^{29}Si (Holzman *et al.* 1956), ^{81}Br (Itoh and Yamagata 1956), ^{87}Rb and ^{133}Cs (Gutowsky and McGarvey 1953), ^{207}Pb (Piette and Weaver 1958), various heavy nuclei (Proctor and Yu 1951).

§ 3. Indirect Spin–Spin Interactions in Steady State Spectra

3.1. *Introduction*

In many cases the number of lines in the nuclear magnetic resonance spectra of liquids is much higher than could be accounted for by the shielding differences discussed in the previous section even when questions of chemical equivalence are not invoked or involved. The effects were discovered in both steady state (Gutowsky, McCall, Slichter and McNeil 1951 a, b, Proctor and Yu 1951) and transient experiments (Hahn and Maxwell 1951, 1952, McNeil *et al.* 1951), but since the steady state experiments are rather easier to analyse we shall discuss the effects from that point of view first, reserving the discussion of transients to § 4.

It is not immediately obvious what is the source of these splittings as it was for the magnetic shielding effects and their exact nature is not simply explained, so we shall first describe a few experimental results and give a summary of the general rules obeyed and a simple explanation before proceeding to a more refined analysis. There is again evidence that the bulk of the effects are molecular. The classic example is ethyl alcohol, but we shall consider some simpler cases first.

For the ion PO_3F^{2-} in aqueous Na_2PO_3F both the ^{31}P and the ^{19}F resonances are doublets with the same separation, 860 c/s, although there is certainly only one ^{31}P and one ^{19}F nucleus in the ion, so there is no question of chemical shift. (The field splittings would be in the ratio $\gamma_{^{19}F}/\gamma_{^{31}P} \simeq 2{\cdot}35$.) We shall find that such splittings are more usefully described in terms of frequency (i.e. energy). The ^{19}F doublet in $POCl_2F$, a similar case, is shown in figure 13.

The proton and deuteron absorption lines of gaseous hydrogen deuteride are shown in figure 14, where we note that the *proton* spectrum is a triplet with three lines of equal integrated intensity, and the deuteron spectrum is a doublet with the same spacing (43 c/s) between each line (the traces in figure 14 show the field splittings superimposed; they are in the ratio $\gamma_H/\gamma_D \simeq 6{\cdot}5$). It would require a quite unreasonable association between molecules to give such an effect. The interaction is clearly internal to the molecule.

An obvious clue to this behaviour is that in the first example (figure 13) the ^{19}F nucleus is subjected to a small magnetic field which depends on the two

possible, almost equally probable, orientations of the ^{31}P nucleus. Again, in HD (figure 14) the proton appears to react to the three possible Zeeman states of the deuteron (which has $I = 1$) and conversely the deuteron responds to the two states of the proton, but in both cases the interaction is entirely mutual since the splitting energy is exactly the same.

We summarize the observed properties of these multiplet splittings before discussing their origin. The symbol J will be used to designate this interaction.

Figure 13. The doublet ^{19}F resonance in $POCl_2F$. The splitting is 1250 c/s (Gutowsky *et al.* 1953).

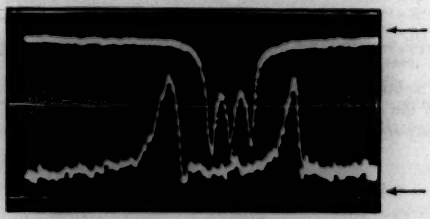


Figure 14. The triplet proton and doublet deuteron signals in gaseous HD artificially superimposed (Wimett 1953).

(1) Splittings cannot be assigned to interactions between structurally equivalent nuclei. Equivalent nuclei may be defined as nuclei in positions which can be brought into coincidence by a symmetry operation on the molecule. Symmetry must include the isotopic nature of the magnetic nuclei. In particular, this means that equivalent nuclei are not only identical nuclei but ones with the same shielding field. Equivalent nuclei have identical couplings of all kinds with all other nuclei in the molecule. The symmetry definition is not entirely satisfactory and the statement just made is more exact since nuclei may still be magnetically inequivalent. For example, in the plane 'molecule' $\begin{matrix} A_1 & & B_1 \\ & \times & \\ A_2 & & B_2 \end{matrix}$ we may have $J_{A_1B_1} \neq J_{A_1B_2}$, hence A_1 and A_2 are not identical because they can tell B_1 from B_2. This causes a relaxation of our rules given below in the sense that one ought to be able to determine from the spectrum all four different coupling constants $J_{A_1B_1}$, $J_{A_1B_2}$, $J_{A_1A_2}$ and $J_{B_1B_2}$. Such a case is 1,1-difluoroethylene

($H_2C{=}CF_2$) which is analysed by McConnell *et al.* (1955). J coupling can also be observed between nuclei which 'appear to be equivalent', for example as a result of internal rotation (Shoolery and Crawford 1957). Resonance lines in all compounds containing only equivalent nuclei (possibly together with non-magnetic nuclei) are single (e.g. H_2, UF_6, etc.).

(2) The splitting is always a mutual effect between non-equivalent nuclei. Its magnitude generally increases with the atomic number of both nuclei. (In the series shown below the magnetic moments generally fall with increasing atomic number, so this does not appear to be a vital factor in comparing nuclei of different atomic number.)

Nuclei interacting	Typical splitting value (c/s)	Observed in
$^{1}H \times {}^{1}H$	4	C_2H_5OH
$^{1}H \times {}^{11}B$	87	$Al(BH_4)_3$
$^{1}H \times {}^{13}C$	153	CH_3Br
$^{1}H \times {}^{14}N$	90	NH_3
$^{1}H \times {}^{19}F$	7	C_6H_5F
	40	CH_3F
$^{1}H \times {}^{31}P$	180	PH_3
$^{1}H \times {}^{207}Pb$	124	$(C_2H_5)_4Pb$
$^{19}F \times {}^{19}F$	75	BrF_5
$^{19}F \times {}^{13}C$	600	C_6H_5F
$^{19}F \times {}^{31}P$	610	PF_3
$^{203}Tl \times {}^{205}Tl$	15000	Thallium metal

(3) The splitting is independent of the applied field (provided it is much smaller than the chemical shift of the nuclei concerned at that field). Examples are (Gutowsky *et al.* 1953):

		Resonance splitting (c/s)		
Substance	Field (G)	^{1}H	^{19}F	^{31}P
CH_3OPF_2	4180	—	1310	1300
($^{19}F \times {}^{31}P$)	6365	—	1280	1290
$HPO(OH)_2$	4180	705	—	707
($^{1}H \times {}^{31}P$)	6365	705	—	713

and for HPF_6 ($^{19}F \times {}^{31}P$) (Quinn and Brown 1953, Gutowsky *et al.* 1953, Roux 1957):

Field (G)	6365	4180	165	35	15
Splitting (c/s)	712	712	708	712	710

Constancy of the $^{1}H \times {}^{31}P$ splitting down to $H_0 = 2$ gauss has been observed in NaH_2PO_2 (Rocard 1957). In contrast the diamagnetic shielding difference in frequency units is directly proportional to the applied field. Even at $H_0 = 2$ gauss the 'chemical shift' between ^{1}H and ^{31}P is still 5066 c/s, i.e. much greater than 710 c/s. See also ^{121}Sb and ^{123}Sb resonances in aqueous solution of $NaSbF_6$ (Proctor and Yu 1951, Dharmatti and Weaver 1952).

(4) There is no obvious and general relation between multiplet splitting and magnetic shielding (chemical shift).

(5) If two types of nuclei in the same compound split each other the field splitting is inversely proportional to the magnetic moment if the nuclei have the same spin, or to the gyromagnetic ratio in general, i.e. the frequency or energy splitting is the same as already mentioned in the preamble (e.g. HD, SbF_6^-, etc.).

(6) If there is only one nucleus causing the splitting the number of lines is the $2I+1$ value of that nucleus. The lines are of equal intensity and the frequency splitting between any pair of multiplet components is equal if the splitting is smaller than the shielding difference in frequency units, e.g. HD in figure 15. In practice the splitting is always smaller than the shielding difference unless the nuclei are identical, but of course then they must be inequivalent. For convenience we shall call the multiplet splitting J and the shielding difference at a given field δ, so that this condition is $J \ll \delta$. We shall take both J and δ to be in cycles per second. In the literature they are sometimes given in radians per second.

(7) A resonance line can be split by a set of equivalent nuclei and one then gets $2I_t+1$ lines where $I_t = \Sigma_i I_i$ and the intensity of any one line is proportional to the number of ways of getting $m_t = \Sigma m_i$ (i.e. the binomial coefficients) and this does not depend on the number of equivalent nuclei in the set being split provided $J \ll \delta$; for example, the methyl proton resonance in ethyl alcohol (CH_3CH_2OH) is a 1,2,1 triplet due to interaction with the methylene group protons for which $I_1 = I_2 = \frac{1}{2}$ and $m_t = -1,0,0,1$.

(8) In a number of cases the splitting expected according to the above rules is not found but satisfactory explanations are available in all cases, for instance, nuclei carrying large electric quadrupole moments do not usually cause splitting in the liquid because of the disorienting effect of the quadrupole relaxation. More cases, including those for which J is not much smaller than δ, will be discussed below.

The above effects are readily explained in a naïve manner by assuming that there is an interaction between nuclei which depends on their orientation and that the main field H_0 plays no direct part, only serving to give the axis of quantization and the basic Zeeman levels. This is consistent with points (3), (4), (5) and (6). Point (1) is not readily understood on this simple picture.

3.2. *Phenomenological Explanation*

A phenomenological explanation is obtained if one includes in the Hamiltonian terms of the form $-hJ_{ij}\mathbf{I}_i.\mathbf{I}_j$ so that the problem is then to explain a term of this form and the magnitude of the J_{ij}. Before doing this let us consider a system containing two sets of non-equivalent nuclei which do not interact. Equation (4′) then becomes

$$\mathscr{H}_m = -\hbar\gamma_1 H_0(1-\sigma_1) I_{1z} - \hbar\gamma_2 H_0(1-\sigma_2) I_{2z}, \qquad \ldots\ldots(27)$$

or if ν_m is the mean of the two resonant frequencies and δ is their difference frequency in the field H_0 we can put (if $\sigma_1 > \sigma_2$, δ is positive)

$$\mathscr{H}_m = -h(\nu_m - \tfrac{1}{2}\delta) I_{1z} - h(\nu_m + \tfrac{1}{2}\delta) I_{2z}. \qquad \ldots\ldots(27')$$

Hence the energy levels are

$$E_{m_1,m_2}/h = -(\nu_m - \tfrac{1}{2}\delta) m_1 - (\nu_m + \tfrac{1}{2}\delta) m_2 \qquad \ldots\ldots(28)$$

and resonance of (1) (or (2)) corresponds to $|\Delta m_1|$ (or $|\Delta m_2|$) = 1 and Δm_2 (or Δm_1) = 0, thus giving the energy levels and transitions on the left of figure 15 for $I_1 = I_2 = \frac{1}{2}$. Thus the observed spectrum is as in the lower part of figure 16 and the resonance lines are (somewhat artificially) doubly degenerate. Consider now the Hamiltonian

$$\mathscr{H} = -h(\nu_m - \tfrac{1}{2}\delta) I_{1z} - h(\nu_m + \tfrac{1}{2}\delta) I_{2z} - hJ\mathbf{I}_1.\mathbf{I}_2 \qquad \ldots\ldots(29)$$

where J is a coupling constant in c/s.

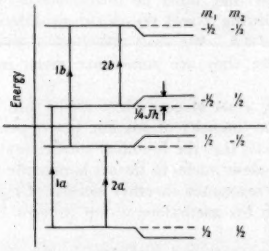


Figure 15.

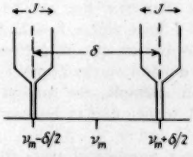


Figure 16.

If $J \ll \delta$ the last term is a small perturbation and the non-diagonal elements in the Hamiltonian matrix may be neglected, i.e. we may replace $\mathbf{I}_1.\mathbf{I}_2$ by $I_{1z}I_{2z}$ and retain m_1 and m_2 as good quantum numbers with the transition probabilities unchanged ; hence

$$E_{m_1, m_2, J}/h = -(\nu_m - \tfrac{1}{2}\delta) m_1 - (\nu_m + \tfrac{1}{2}\delta) m_2 - Jm_1 m_2, \qquad \ldots\ldots(30)$$

and so for excitation of nuclei (1) or (2)

$$\Delta E_1/h = \nu_m - \tfrac{1}{2}\delta + Jm_2 \quad \text{or} \quad \Delta E_2/h = \nu_m + \tfrac{1}{2}\delta + Jm_1.$$

The energy levels are now as on the right of figure 15, that is to say, the perturbation has lifted the degeneracy and gives equal splitting of the two resonances by J c/s (figure 16). If the second nucleus has say $I_2 = 1$, then $m_2 = -1, 0, 1$ and nucleus (1) line is split into a 1,1,1 triplet (e.g. ^{1}H in HD). The observed spectrum is independent of the sign of J.

If there is more than one equivalent nucleus, of type (2) say, we have extra terms in the Hamiltonian like $-hJ_{12}\mathbf{I}_1.\mathbf{I}_{2i}$ where the J_{12} must be independent of i, so that we have a term $hJ_{12}\mathbf{I}_1.\mathbf{I}_{2t}$ where $\mathbf{I}_{2t} = \Sigma_i \mathbf{I}_{2i}$. There must also be terms like $\mathscr{H}_{ij} = -hJ_2\mathbf{I}_{2i}.\mathbf{I}_{2j}$, since there is no reason why equivalent nuclei should not interact with each other or even equally with each other. It can, however, be shown by an elegant theorem (Gutowsky, McCall and Slichter 1953) that the time dependence of the wave functions is independent of $\mathscr{H}_{ij}$ and so in nuclear resonance experiments such interactions between equivalent spins may be ignored.

However, the J coupling between equivalent nuclei can sometimes be determined by making them 'slightly inequivalent', as discussed in § 3.3. Hence the most general interaction between two sets A and B of equivalent spins is of the form

$$-hJ_{AB}\mathbf{I}_A.\mathbf{I}_B \quad \text{where} \quad \mathbf{I}_A = \sum_{i=A} I_i.$$

Thus using the same approximation $J_{AB} \ll \delta_{AB}$, we would have

$$\Delta E_A/h = \nu_A + J_{AB}M_B, \quad \Delta E_B/h = \nu_B + J_{AB}M_A, \qquad \ldots\ldots(31)$$

which gives a multiplet splitting where the number and relative intensities of the lines for one nucleus depends on the number of ways of having the various values of M for the *other* set. The small splittings are conveniently measured by the 'sideband technique' (Arnold and Packard 1951, Anderson 1956 a). The rule is

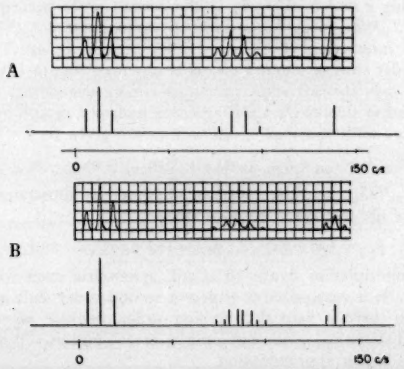


Figure 17. Ethyl alcohol proton spectrum at medium resolution. Groups of lines from left to right are from the methyl, methylene and hydroxyl protons respectively. A, trace and theoretical spectrum for 'wet' alcohol, and B, for 'dry', showing the effect of hydroxyl proton exchange (Arnold 1956).

very simple in application, for example, in 'wet' ethyl alcohol (see figure 17(*a*)) the methyl spectrum is a 1,2,1 triplet corresponding to the states $\alpha\alpha$, $\alpha\beta$, $\beta\alpha$, $\beta\beta$ of the methylene protons noting that, for both $\alpha\beta$ and $\beta\alpha$, $M = 0$. Again the methylene protons signal should be a quartet 1,3,3,1 (binomial coefficients of '3') corresponding to the states of CH_3, $\alpha\alpha\alpha$; $\alpha\alpha\beta$, $\alpha\beta\alpha$, $\beta\alpha\alpha$; etc., i.e. $M = \frac{3}{2}$ once, $\frac{1}{2}$ three times, $-\frac{1}{2}$ three times, and $-\frac{3}{2}$ once. ($J \ll \delta$, since at 30·5 Mc/s, $\delta_{CH_3-CH_2} = 75$ c/s and $J_{CH_3-CH_2} = 7{\cdot}1$ c/s.)

More complicated cases arise. In ethyl alcohol one would expect an interaction between the methylene (CH_2) and hydroxyl (OH) protons with a different J value so that the CH_2 spectrum (if $J \ll \delta$) would be doubly split, since there is now an extra term $-hJ_{BC}\mathbf{I}_B.\mathbf{I}_C$ in the Hamiltonian and $J_{BC}M_C$ in equation (31).

This gives an eight-line spectrum which, since it happens that $J_{CH_3-OH} \simeq \frac{1}{2} J_{CH_3-CH_2}$ (actually $J_{CH_3-OH} = 4{\cdot}8$ c/s), is of the form 1,1 ; 3,3 ; 3,3 ; 1,1, as in figure 17(*b*) for 'dry' alcohol. (The difference for 'wet' alcohol is due to exchange of hydroxyl protons, which is discussed later.) Since it is reasonable to suppose that $J_{CH_3-OH} \ll J_{CH_2-OH}$, the OH resonance should be a 1,2,1 triplet, as observed. Obviously ethyl alcohol constitutes a rather simple case and many spectra where J values are comparable in magnitude and not much less (or much greater) than δ will be more complex. Special methods have been developed to deal with such situations and these are discussed in § 3.6. Under even higher resolution (higher main field homogeneity) further splittings appear which were not foreseen, due to the use of the approximate Hamiltonian with $I_{1z} I_{2z}$ instead of $\mathbf{I}_1 . \mathbf{I}_2$. In ethyl alcohol $J/\delta \simeq \frac{1}{10}$, so that it is worth while, as in many other cases, to continue the perturbation calculation to second order. Since the second-order perturbation energy is of order [(matrix element of M_x)^2/energy difference] and the only operator connecting a and b is $hJ_{ab} \mathbf{I}_a . \mathbf{I}_b$, the second-order perturbation energy is approximately $h^2 J_{ab}^2/h(\nu_a - \nu_b)$, i.e. a shift of approximately J_{ab}^2/δ_{ab} c/s. Hence, for example, the interaction of CH_3 and CH_2 protons in ethyl alcohol should show a second-order shift of about 1 c/s, as is observed (figure 19). The second-order interaction may also lift some of the remaining degeneracy. For instance, if we consider in ethyl alcohol the splitting of the hydroxyl proton by the methylene protons, the energy level spacings of the former are given by

$$\Delta E_{OH}/h = \nu_{OH} + J_{CH_2-OH}\, m_{CH_2} + \Delta[\mathbf{m}_{CH_2}^2 + \mathbf{m}_{CH_2} - m_{CH_2}^2]$$

where $\Delta = J_{CH_2-OH}^2/2\delta_{CH_2-OH}$, the meaning of which is illustrated in figure 18, where the states of the methylene protons are

$$\phi_{1-1} = \beta\beta, \quad \phi_{10} = (\alpha\beta + \beta\alpha)/\sqrt{2}, \quad \phi_{00} = (\alpha\beta - \beta\alpha)/\sqrt{2} \quad \text{and} \quad \phi_{11} = \alpha\alpha.$$

Notice that the multiplet in figure 18 is still symmetric even with the second-order correction. It is reasonable to expect a second-order shift arising from the state ϕ_{10} although there is zero shift in first order, because $m_z = 0$ but $\mathbf{m} = 1$. No shift is expected for ϕ_{00} since not only is $m = 0$ but $\mathbf{m} = 0$ and the state is non-magnetic to a higher approximation.

Figure 18.

In general the second-order shifts are more complex since for two sets A and B the m and $\mathbf{m}$ values of *both* A and B are now relevant and the energy levels are in fact

$$E_{m_A, m_A, m_B, m_B}/h = -\nu_A m_A - \nu_B m_B - J_{AB} m_A m_B + J_{AB}^2/2\delta_{AB} \times [m_A(\mathbf{m}_B^2 + \mathbf{m}_B - m_B^2) - m_B(\mathbf{m}_A^2 + \mathbf{m}_A - m_A^2)].$$

The observed and calculated spectra of the methyl protons (and even of the methylene protons) in ethanol shown in figure 19 are readily interpreted on the basis of these simple ideas, apart from some slight intensity changes which are discussed below. The spectrum of ethyl bromide (CH_3CH_2Br) is very similar (Anderson 1956 a, b). For methanol (CH_3OH), see Kivelson and Kivelson (1958). This exact agreement must be regarded as a rather fine vindication of the truth of the assumptions made.

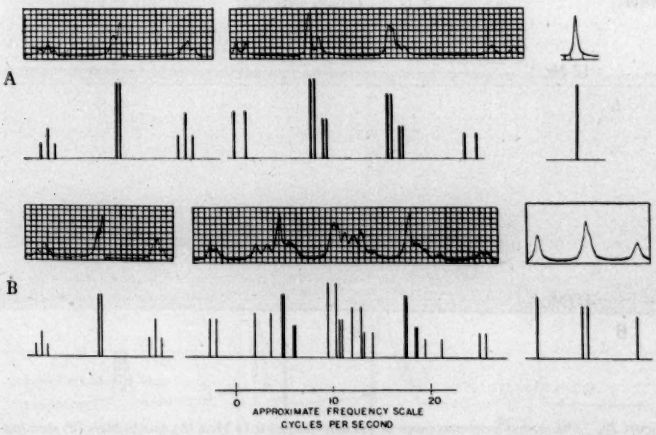


Figure 19. Observed and calculated spectra for ethyl alcohol at high resolution showing second-order splittings, A with and B without hydroxyl proton exchange (Arnold 1956).

Impressive as are the spectra of figure 19, they do not really contain more information than the spectra of figure 17, but merely confirm the interpretation. If one seeks only the δ and J values, which are the physical molecular properties, or even simply to use the spectrum as an analytical tool, there is considerable advantage in minimizing the second-order splittings. Since they depend on J^2/δ and $\delta \propto H_0$, this can be done by increasing the field H_0. An impression of the change in the complexity in a typical case is shown in figure 20 for the wet ethyl alcohol methylene proton signal. (N.B. Field changes are in opposite sense compared with figures 17 and 19.) Spectrum B is a recognizable perturbation of the simple 1 ; 3 ; 3 ; 1 spectrum. If the field homogeneity is maintained one also has the advantage of increased signal strength. These spectra also reveal intensity changes which occur when J is not much smaller than δ ; these are discussed below.

For arbitrary values of J and δ an exact theory must be used and the general behaviour is nicely illustrated by the calculation for two interacting nuclei of spin $\frac{1}{2}$ (Hahn and Maxwell 1952).

The four wave functions, corresponding to the zero-order wave functions $\alpha\alpha, \alpha\beta, \beta\alpha, \beta\beta$ for two non-interacting inequivalent nuclei (i.e. for $J = 0, \delta \neq 0$), are

$$\alpha\alpha, \quad \frac{1}{\sqrt{2(1+R^2)^{1/2}}}[\alpha\beta(1+R)+\beta\alpha(1-R)],$$

$$\frac{1}{\sqrt{2(1+Q^2)^{1/2}}}[\alpha\beta(1+Q)+\beta\alpha(1-Q)], \quad \beta\beta \qquad \ldots\ldots(32)$$

where $$R, Q = [J \pm (J^2+\delta^2)^{1/2}]/\delta.$$

Figure 20. The methylene proton signal in wet ethyl alcohol at 12 Mc/s (A) and 40 Mc/s (B) showing the decrease in second order splitting at the higher field and consequent simplification of the spectrum (Varian).

When $\delta \to 0$, $R \to \infty$ and $Q \to 0$ and these wave functions become the triplet and singlet wave functions $\alpha\alpha, (\alpha\beta-\beta\alpha)/\sqrt{2}, (\alpha\beta+\beta\alpha)/\sqrt{2}, \beta\beta$, and so even when $\delta \neq 0$ we shall call them $\phi_{1\,1}$, $\phi_{0\,0}$, $\phi_{1\,0}$, $\phi_{1\text{-}1}$. The energy levels are

$$E_{1\,1}/h = -\nu_m - \tfrac{1}{4}J, \qquad E_{1\text{-}1}/h = \nu_m - \tfrac{1}{4}J,$$
$$E_{1\,0}/h = \tfrac{1}{4}J - \tfrac{1}{2}(J^2+\delta^2)^{1/2}, \quad E_{0\,0}/h = \tfrac{1}{4}J + \tfrac{1}{2}(J^2+\delta^2)^{1/2}. \qquad \ldots\ldots(33)$$

The transition probabilities for $(0,0) \rightleftharpoons (1,-1)$, $(1,1) \rightleftharpoons (1,0)$, $(1,0) \rightleftharpoons (1,-1)$ and $(1,1) \rightleftharpoons (0,0)$ are in the ratios $1/(1+R^2) : 1/(1+Q^2) : 1/(1+Q^2) : 1/(1+R^2)$. When $J \ll \delta$ the intensities only differ by terms in second order in J/δ since $R, Q \to \pm 1 + J/\delta$ and we readily recover the spectrum 1,1;1,1 with spacing J between doublets as in the approximate theory (figure 16). The next approximation readily gives the second-order shifts and further shows that as J/δ increases the intensity of the inner lines tends to increase at the expense of the outer ones. This is shown in figure 21 for $J/\delta = 0{\cdot}835$ as observed in 2-bromo-5-chlorothiophene (Anderson 1956 a).

The exact result for this case which we shall need later is, for the lines in increasing frequency (if J is positive),

(1) $\nu_m - \frac{1}{2}J - \frac{1}{2}(J^2+\delta^2)^{1/2}$ intensity $\frac{1}{4}[1 - J/(\delta^2+J^2)^{1/2}]$,
(2) $\nu_m + \frac{1}{2}J - \frac{1}{2}(J^2+\delta^2)^{1/2}$ intensity $\frac{1}{4}[1 + J/(\delta^2+J^2)^{1/2}]$,
(3) $\nu_m - \frac{1}{2}J + \frac{1}{2}(J^2+\delta^2)^{1/2}$ intensity as (2),
(4) $\nu_m + \frac{1}{2}J + \frac{1}{2}(J^2+\delta^2)^{1/2}$ intensity as (1). .....(34)

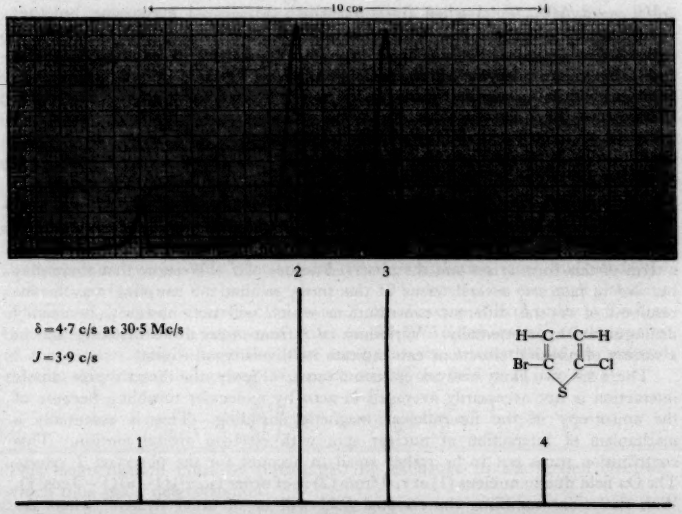


Figure 21. The spectrum of two single inequivalent protons with J coupling and $J \simeq \delta$. 2-bromo-5-chlorothiophene at 30·5 Mc/s. Theoretical spectrum is for $J/\delta = 0{\cdot}835$ using equation (34). (Anderson 1956a).

In the limit when $\delta \to 0$ the outer lines (1) and (4) weaken towards zero with separation J from the centre and the inner lines move to the centre and we get a single line independent of J. Thus J is unobservable when the two nuclei are equivalent, since all three triplet levels involved in transitions are shifted by the same amount (eqn (33)) with $\delta = 0$ and transitions to and from the singlet level are forbidden. This result, which is generally true, arises specifically from the use of $\mathbf{I}_1.\mathbf{I}_2$ rather than $I_{1z}I_{2z}$ in the Hamiltonian. The latter gives a splitting for $\delta \to 0$ into a doublet with separation J in this simple case, and this is certainly incorrect. The actual splitting of the inner line is approximately $\delta^2/2J$ when $\delta \ll J$. This is a specific illustration of the general theorem that indirect coupling between

equivalent nuclei cannot be observed except in the impracticable case when the resonant frequency is comparable with J.

The spectra to be expected in a number of more complicated cases have been computed by Hahn and Maxwell (1952), Banerjee, Das and Saha (1954), Arnold (1956) (see also Bernstein, Pople and Schneider (1957) and Pople, Schneider and Bernstein (1957)). The only new feature arising is that some extra lines appear which would not be predicted from the simple picture when $J \ll \delta$ even allowing for degeneracy. For two independent groups of nuclei A and B the selection rules are $\Delta M_{z\mathrm{A}} = \pm 1$, $\Delta M_{z\mathrm{B}} = 0$ or $\Delta M_{z\mathrm{A}} = 0$, $\Delta M_{z\mathrm{B}} = \pm 1$, i.e. $\Delta M_{zt} = \pm 1$. But if the two groups interact, as with J coupling, transitions corresponding to $\Delta M_{z\mathrm{A}} = +2$, $\Delta M_{z\mathrm{B}} = -1$ when $J = 0$ for which $\Delta M_{zt} = +1$ may occur, but they have intensity of order $(J/\delta)^2$ when $J \ll \delta$. Such lines need special consideration in second-moment calculations since although second-order in intensity they have first-order shifts (truncated Hamiltonian, etc. (Van Vleck 1948)).

Again, if there are, say, two sets of nuclei of the same species containing different numbers of equivalent nuclei, the complete spectra become asymmetric (as in ethyl alcohol) and so the sets may be identified, or in other words, the sign of the difference in shielding may be determined.

3.3. *The Physical Origin of Indirect Spin–Spin Coupling*

All the experimental results are explained for liquids by introducing a term of the form $-hJ_{12}\mathbf{I}_1.\mathbf{I}_2$ into the Hamiltonian. We have therefore to explain how a term of this form arises and the observed values of J. We note that there may be, and in fact are, several terms of this form, so that the coupling may be the resultant of several different contributions which will not, however, be readily distinguished experimentally. Variations in J from molecule to molecule and in a variety of physical situations can indicate its diversity of origin.

There are two likely sources of such a term. Firstly, the direct dipole–dipole interaction is not necessarily averaged to zero by molecular tumbling because of the anisotropy of the internuclear magnetic shielding. This is essentially a mechanism of interaction of nuclear spin with electron orbital motion. This contribution turns out to be rather small to account for the observed J values. The Oz field due to nucleus (1) at r, θ from (2) is of order $(\mu_1/r^3)(1-\sigma)(1-3\cos^2\theta)$. With molecular tumbling the residual field will be of order $\mu_1\Delta\sigma/r^3$ where $\Delta\sigma$ is the anisotropy of the magnetic shielding at the second nucleus and we ignore for the moment the fact that $\Delta\sigma$ really refers to a uniform external field and not to a dipole field. The interaction energy is therefore of order $(\mu_1/r^3)\Delta\sigma\mu_2$, i.e. it is the direct interaction reduced by a factor $\Delta\sigma$ which might be 10^{-5} for protons. This interaction is therefore almost as small as the direct interaction attenuated by molecular tumbling which we have shown to be usually negligible in the cases under consideration. In fact, although the contribution is small, it is not negligible. A second source of interaction was pointed out by Ramsey and Purcell (1952) in that there can be an indirect coupling via the electron *spin*. If a nuclear spin (1) is in a particular state, for example a proton in α, then owing to a nuclear–electronic spin interaction there may be a slightly larger probability that the spin of that electron will lie antiparallel than parallel to it when near, i.e. that it will be in the β state (or vice versa depending on the interaction).

However, this electron is one of a coupled pair so that the other electron of this (molecular) orbital in the bond connecting the two nuclei (1) and (2) will tend to be in the α state. Hence the nucleus (2) by the same token (or possibly by another) will tend to be in the β (or the α) state. In this way the nuclei interact via the electron spins in the bonding electrons and the effect is obviously quite mutual and can moreover be at least partially independent of molecular orientation, so that an interaction of the form $\mathbf{I}_1.\mathbf{I}_2$ is feasible. In this simple analysis there is therefore a tendency for the nuclei to be antiparallel. With the convention of equation (29)† J is negative in this case since then the indirect coupling lowers the magnetic energy for antiparallel nuclei (see figure 15, which is drawn assuming J positive, i.e. nuclei tend to be parallel). However, it is possible to see even with this simple picture that J can have either sign. The nuclei may interact with the spin of one electron, in which case they would tend to be parallel, i.e. J positive; however, this interaction is likely to be small, like the orbital one discussed below. Not very serious differences arise from a difference in sign of J since the nuclear wave functions do not overlap and there are no exchange effects of electrostatic origin as for electrons. The determination of the sign of J experimentally is quite difficult.

Electron spin effects are now to be included because the field of a nucleus varies over the molecule and so does not affect equally the two electron spins of a coupled pair as does a uniform external field.

An order of magnitude for J is readily found (Gutowsky *et al.* 1953). Such an interaction corresponds to a second-order perturbation since one needs the excited triplet states of the bonding electrons. The interaction energy of a nucleus of moment μ_1 and an electron of moment β at a distance r_1 is of order $\mu_1\beta/r_1^3$, so that the perturbation energy is of order $(\mu_1\beta/r_1^3)^2/\Delta E_s$ where ΔE_s is the energy of the excited state. Regarding this energy as the interaction of an induced electron spin magnetic moment M_{1s} produced by the field of μ_1, we have

$$M_{1s}\frac{\mu_1}{r_1^3} \simeq \left(\frac{\mu_1\beta}{r_1^3}\right)^2 \Big/ \Delta E_s, \quad \text{or} \quad M_{1s} \simeq \frac{\mu_1\beta^2}{\Delta E_s}\left\langle\frac{1}{r_1^3}\right\rangle.$$

This magnetization must be transmitted by exchange to the second electron, which thus has a magnetization

$$M_{2s} \simeq -\frac{\mu_1\beta^2}{\Delta E_s}\left\langle\frac{1}{r_1^3}\right\rangle$$

and produces a field at the nucleus (2) and the interaction energy is of order

$$M_{2s}\mu_2\left\langle\frac{1}{r_2^3}\right\rangle = \frac{\mu_1\mu_2\beta^2}{\Delta E_s}\left\langle\frac{1}{r_1^3}\right\rangle\left\langle\frac{1}{r_2^3}\right\rangle, \qquad \ldots\ldots(35)$$

which is bilinear in μ_1 and μ_2 (and hence in I_1 and I_2).

The orbital magnetization produced by nucleus (1) in the (1) electron is similarly

$$M_{10} \simeq \frac{\mu_1\beta^2}{\Delta E_0}\left\langle\frac{1}{r_1^3}\right\rangle,$$

† The opposite convention is equally common.

but this orbital moment is not transferred by 'exchange' but interacts directly with nucleus (2) which may be R_{12} away, giving an interaction energy of order

$$\frac{M_{10}\mu_2}{R_{12}^3} \simeq \frac{\mu_1\mu_1\beta^2}{\Delta E_0}\left\langle\frac{1}{r_1^3}\right\rangle\left\langle\frac{1}{R_{12}^3}\right\rangle. \qquad \ldots\ldots(36)$$

Thus the orbital interaction is smaller than the spin interaction by the factor $(1/R_{12}^3)/\langle 1/r_2^3\rangle E_0/\Delta E_s$, which may be ten or more. It is therefore to be expected that the electron spin interaction provides the larger contribution in most cases.

Comparing equation (35) with $hJ\mathbf{I}_1.\mathbf{I}_2$, we see that for the spin interaction

$$J \simeq \frac{\mu_1\mu_2\beta^2}{\Delta E_s}\left\langle\frac{1}{r_1^3}\right\rangle\left\langle\frac{1}{r_2^3}\right\rangle\frac{1}{hI_1I_2} = \frac{\gamma_1\gamma_2\beta^2h}{4\pi^2\Delta E_s}\left\langle\frac{1}{r_1^3}\right\rangle\left\langle\frac{1}{r_2^3}\right\rangle.$$

Suppose we have two protons and assume that $\langle 1/r_1^3\rangle = \langle 1/r_2^3\rangle = 1/a_0^3$ and $\Delta E_s = 10$ ev, then this formula gives $J \simeq 40$ c/s, which is of the correct order of magnitude.

In writing the Hamiltonian of the molecule we are now interested in terms which lead directly or in combination to energy terms bilinear in I_1 and I_2, i.e. $W = W_0 - hJ\mathbf{I}_1.\mathbf{I}_2$. In the Hamiltonian, equation (20) of § 2·2, we noted a term NN which has the desired form but, although this is non-zero in first-order perturbation, it will be small, since whenever r_i is small, r_j is large and vice versa. Further, since evidently inner electrons are important and these will be in closed shells, it will be very seriously reduced by the molecular tumbling.

As for the chemical shift calculation, we could now find further terms bilinear in μ_1 and μ_2 by a second-order perturbation calculation, and this is at the basis of the method of Ramsey (1953) discussed by Abragam (1957). The terms to be considered are ON and two terms which were not included in the Hamiltonian (20) which are

$$\mathrm{NS} = -2\beta[\mathbf{S}_k.\boldsymbol{\mu}_N r_{kN}^{-3} - 3(\mathbf{S}_k.\mathbf{r}_{kN})(\boldsymbol{\mu}_N.\mathbf{r}_{kN})r_{kN}^{-5}]$$

the dipolar interaction between nuclear, N, and electronic, k, spins, and

$$\mathrm{SF} = \frac{16\pi}{3}\beta\mathbf{S}_k.\boldsymbol{\mu}_N\delta(\mathbf{r}_{kN})$$

the Fermi (1930) type electron-spin–nuclear-spin interaction. These terms are of no present interest in a first-order perturbation.

The terms ON, NS and SF combined with themselves in a second-order perturbation give rise to interactions of the correct form. (Cross terms vanish because of molecular tumbling in the liquid.) The importance of the different terms may be expected to vary in different cases, NS and SF being probably always the most important.

The general expression for J obtained by Ramsey (1953) in this way is almost impracticable. For HD, which is the simplest example (H_2 has no observable J coupling, but J_{HH} can be deduced from J_{HD}) it is found that over 90 per cent of the interaction arises from SF, because of the dominant s character of the electrons and that

$$J = \frac{64\beta^2 h\gamma_H\gamma_D}{9\Delta E}|\phi|^2_{H,D},$$

where ΔE is a mean excitation energy and $|\phi|^2_{H,D}$ is the probability of one electron being at the proton and the other at the deuteron. The numerical result is $J \simeq 70$ c/s with a guessed value of ΔE (Ramsey and Purcell 1952). The measured value is $43 \pm 0{\cdot}5$ c/s (Carr and Purcell 1952, Wimett 1953).

We shall not discuss such calculations in detail since they are of little value except for giving an order of magnitude and for fairly general considerations since the energy values ΔE are not available. McConnell (1956) obtains an interesting relation between indirect interaction for protons and bond order in hydrocarbons. For J coupling of aromatic protons, see McConnell (1957 b). Satisfactory order of magnitude results for fluorine interactions have been obtained (Gutowsky *et al.* 1953).

As for the magnetic shielding calculation an obvious alternative is a variational calculation and this has met with some success. Stephen (1957 b) used perturbed wave functions of the form $\phi = \phi_0(1+\mathbf{I}_A.\mathbf{f}_A+\mathbf{I}_B.\mathbf{f}_B)$ where the $\mathbf{f}$ are unknown functions and the J value is found by minimizing the energy $W = W_0 - hJ\mathbf{I}_A.\mathbf{I}_B$, although it would appear that the total energy should be minimized. The calculations are rather similar to those for diamagnetic shielding and the various contributions to J can be calculated using valence bond or molecular orbital wave functions. The calculations have been carried out for HD with the following results (Stephen 1957 b). J has the same sign for both contributions (Stephen, private communication).

	Valence bond	Molecular orbital
J due to ON	1·46	2·08
J due to SF and NS	47·68	47·49
Total J (c/s)	49·1	49·67
Experiment (c/s)	$43 \pm 0{\cdot}5$	

The result is almost independent of the type of wave function used since the principal contribution to J is from the SF term and this depends mainly on the wave function near the nucleus which is similar in the two cases. This state of affairs is probably not generally valid outside proton–proton indirect interactions. The sign of the various contributions to J is relevant. Ishiguro (1958) has calculated J_{HD} as a function of internuclear distance which permits an allowance for internal vibration. The actual numerical result is, however, only approximate, giving 35·2 c/s.

Pople (1958 c) calculates by an approximate method the orbital contribution to the proton–proton and proton–fluorine indirect coupling in a number of molecules. The contribution is always much less than the coupling via electron spin but is by no means negligible. It may have either sign.

McConnell (1956) has calculated the J values for interactions through several chemical bonds (see also Aihara 1957). In this case interactions via electron orbital motion may be comparable with the spin interaction because of electron exchange.

The J coupling may connect nuclei separated by as many as four chemical bonds (Saika and Gutowsky 1956) so that in general spectra may be expected to

be extremely complex. On the other hand, there are often a number of simplifications. For instance, in p-fluoro-ethyl benzene (Baker 1955) the ^{19}F spectrum is a 1 : 4 : 6 : 4 : 1 with $J \simeq 7$ c/s apparently indicating interaction with four equivalent protons, presumably only the ring ones $2H_a$ and $2H_b$. Thus there is an electronic coupling round the benzene ring which does not apparently flow in the ethyl group, which therefore may be treated as an almost independent system. As is consistent with this the ring proton signal is a doublet with $J = 6{\cdot}6$ c/s. It appears therefore that the two types of ring protons do not have a measurable chemical shift at the resolution attained and are almost equivalent. However, H_a and H_b probably have different J couplings with the fluorine nucleus, and in fact the J value observed is $J_{av} = \frac{1}{2}(J_{H_a} + J_{H_b})$. This happens when $J \gg \delta$ and this situation is further discussed in § 4.2.

An example of a satisfactory detailed analysis of a more complex case is that of naphthalene (approximately A_2B_2) where three J values are required (Bernstein, Schneider and Pople 1957, Bernstein, Pople and Schneider 1957, see also Richards and Schaeffer 1958 a, b).

In one case the multiplet splitting has been used to determine a nuclear spin, i.e. that of ^{29}Si is one half (Williams *et al.* 1954), in a manner similar to that which has been widely exploited in electron resonance.

The most intense signals are obtained from protons and because of their widespread occurrence in chemical compounds most work has been done on this nucleus. The next easiest are the ^{19}F and ^{31}P nuclei.

It is a source of regret (but also of simplification) in studying organic molecules that ^{12}C has zero magnetic moment. The 1·1% natural abundant ^{13}C nucleus is not easy to observe directly (but see Lauterbur (1957, 1958) and Holm (1957)). However, the indirect interaction of a ^{13}C nucleus ($I = \frac{1}{2}$) with the protons may betray its presence by weak satellite lines on the main proton signal. The resonance of hydrogen atoms directly attached to a ^{13}C nucleus consists of a widely spaced weak doublet with $J(^{13}C^{1}H) = 120$–200 c/s (Lauterbur 1957, Holm 1957) symmetrically placed about the stronger ^{12}CH resonance, as in figure 22(*a*), which shows only one line of a doublet. We readily understand by use of the simple rules of § 3.1 that the splitting is still a doublet if ^{13}C is coupled to more than one equivalent proton, as in dichloromethane (figure 22(*a*)). When the molecule has adjacent groups with hydrogen atoms which can interact with those attached to the ^{13}C nucleus, one observes a structure in the satellites even if the protons are otherwise equivalent, as for example in trans-dichloroethylene shown in figure 22(*b*) (Cohen, Sheppard and Turner 1958). This is a case of double splitting similar to that already encountered in ethyl alcohol in that H_a is split by ^{13}C ($J \sim 100$ c/s) and each line is then further split into doublet by H_b with $J_{ab} = 12{\cdot}2$ c/s. In the more common molecule with two ^{12}C nuclei, H_a and H_b are equivalent and J_{ab} is not observable. The equivalence of H_a and H_b is removed by the lower symmetry of the $^{13}C^{13}C$ molecule, whereas the molecule is ' chemically ' unaffected, i.e. its electron distribution is almost unchanged. This is therefore a neat method of exposing the J coupling between ' equivalent ' nuclei, which is rather reminiscent

of isotopic substitution methods in infra-red or microwave spectroscopy. Couplings between equivalent protons turn out to be similar in magnitude to coupling between non-equivalent but identical nuclei, as would be expected. More complicated cases are explained similarly. 1-2 dichloro $^{13}C^{12}C$ ethane gives a pair of proton signal satellites with a 1,2,1 structure, as expected. A similar method of measuring J type coupling between equivalent protons is by deuteration. The proton resonance in CH_3D is a 1,1,1 triplet with $J_{HD} = 1{\cdot}9 \pm 0{\cdot}1$ c/s, hence

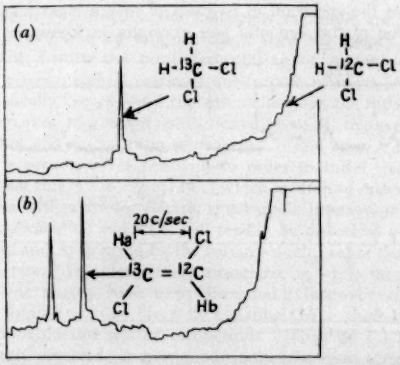


Figure 22. Proton resonance showing one satellite due to ^{13}CH indirect coupling: (*a*) dichloromethane (CH_2Cl_2) in which the two protons are equivalent, (*b*) trans-dichloroethylene (CH_2ClCH_2Cl) showing the doublet due to J coupling between H_a and H_b (Cohen, Sheppard and Turner 1958).

$J_{HH} = (\gamma_H/\gamma_D)(1{\cdot}9 \pm 0{\cdot}1) = 12{\cdot}4 \pm 0{\cdot}6$ c/s. This compares favourably with a theoretical value for the comparatively simple methane molecule of $10{\cdot}4 \pm 1{\cdot}0$ c/s (Karplus *et al.* 1957).

Theory is at present behind experiment in that only rather crude quantitative calculations of J values have been made.

Valuable general conclusions are often forthcoming from the experiments. The results referred to above, for instance, suggest that in substituted ethylenes the indirect coupling is strongest between protons which are trans with respect to each other and this indicates interaction through bonds rather than through space (Cohen *et al.* 1958).

In even moderately complex molecules the number of shifts and possible couplings is so high that some sort of systematic treatment must be used to deal with the complexity.

A notation has been introduced (Bernstein, Pople and Schneider 1957) in which a molecule having $n, m, \ldots$ sets of nuclei with the same shielding all of the same species and $p, q, \ldots$ sets of another species or of the same species but shifted to an extent that $\delta \gg$ all J, is described as an $A_n B_m \ldots X_p Y_q \ldots$ spectrum.

The question of the sign of J is a difficult one. For the Hamiltonian

$$\mathscr{H} = -h\sum_i (\nu_0 + \delta_i) I_{zi} - h\sum_{i<j} J_{ij}\,\mathbf{I}_i.\mathbf{I}_j$$

the spectrum is not affected by the reversal of the sign of all the J_{ij} and so their absolute signs cannot be determined experimentally. Anderson (1956 a) points out that the relative signs are observable if there are at least three interacting groups of nuclei and if second-order splittings are observable (see also Williams and Gutowsky 1956 and Gutowsky *et al.* 1957). We have already noted the symmetry of the spectrum of two sets of inequivalent nuclei which prevents the determination of the sign of J (see examples in Bernstein, Pople and Schneider

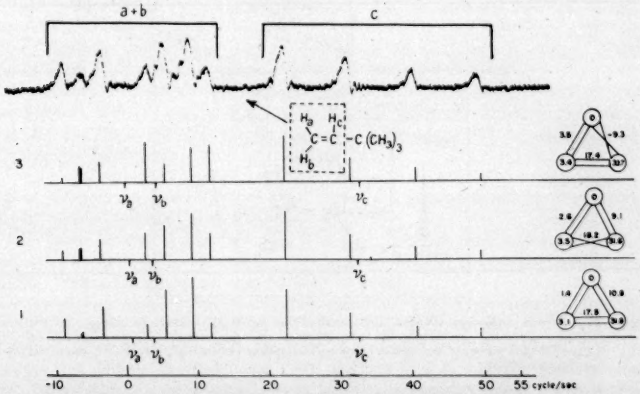


Figure 23. Spectrum of vinyl protons in 3, 3-dimethyl 1 butene at 31·6 Mc/s. (Methyl proton signal is off diagram to left.) Protons H_a and H_b are well shifted from H_c but not much from each other. The theoretical spectra are for three of four possible relative signs of the three J values. Note. The c signal can be regarded as a perturbed 1, 2, 1 due to the two protons a and b. The fit is best for 1, i.e. all J have same sign (Alexander 1958).

1957). The addition of a third set gives sufficient asymmetry to allow the determination of relative signs of couplings, although experimental difficulties tend to intervene. Relative signs of ^1H–^{19}F indirect spin–spin interactions have been found by this method (Williams and Gutowsky 1956). Relative signs can be determined by trial and error provided one has a good idea of the δ and J values. An example is the vinyl ($CH_2{=}CH{-}$) proton signal in 3,3-dimethyl 1-butene [$CH_2{=}CH{-}C(CH_3)_3$] (Alexander 1958) shown in figure 23 (*ABC* spectrum with $B \simeq C$). All three couplings are thought to have the same sign, spectrum 1 in figure 23, although the difference in the spectra for the three possibilities (the fourth possibility is readily excluded) is not great.

Theory suggests that the indirect coupling for protons is usually negative for the convention of equation (29) (McConnell 1956). Thus the signs of other

coupling constants can sometimes be determined (Williams and Gutowsky 1956). However, Alexander (1958) finds for the vinyl (CH_2=CH−) group in 1-butene (CH_2=CH−CH_2CH_3) that of the three proton couplings one has a sign different from the other two. Both signs occur for couplings round the benzene ring (Gutowsky 1958).

A review of the more chemical aspects of indirect spin–spin coupling has been given by Gutowsky (1958).

3.4. *Missing Multiplets*

As mentioned in point (8) of § 3.1 the multiplets expected are sometimes missing. If a nucleus A is J coupled to a nucleus B then any agency disturbing the spin state of B will disturb the multiplet splitting of A, and vice versa. A simple case of this is when spin B carries a quadrupole moment ($J \gg 1$) when quadrupole relaxation usually causes rapid transitions between the different states of B. The spin A then sees an average spin orientation of B, the average value of the interaction is zero and no splitting is observed. The spin A must see a state of B for at least a time $\tau_c = 1/J_{AB}$ seconds in order to suffer clear splitting of its energy levels, so that this is the critical time which is often of order $\frac{1}{10}$ second for protons. A nucleus with a strong electric quadrupole interaction such as is usually the case for chlorine (^{35}Cl and ^{37}Cl) will readily be disturbed at a rate far exceeding this in a liquid and so splitting by the halogen nuclei, other than fluorine, in a molecule will be rare. In intermediate situations, $\tau_c \sim 1/J$, the multiplets are broadened. The line shapes have been discussed (Gutowsky *et al.* 1953, Pople 1958 b) but are similar to those given by equation (26). Such an effect is particularly clear in ammonia and related compounds. The ^{14}N 1,3,3,1 quartet in dry ammonia is poorly resolved, due to quadrupole relaxation (Ogg and Ray 1955, 1957 a) (it collapses to a single line on addition of a trace of water). On the other hand, the NH_4^+ ion, in which the ^{14}N nucleus is not subject to an electric field gradient, has a sharp spectrum. Similarly, $CH_3NH_3^+$, $(CH_3)_2NH_2^+$ and $(CH_3)_3NH^+$ have quadrupole-broadened ^{14}N spectra, but for $(CH_3)_4N^+$ it is sharp (this is in acid solution to suppress hydrolysis exchange). Intermediate cases are not common, one usually observes clear multiplets or no multiplets.

For much the same reason multiplet structure is absent when the second nucleus is being exchanged sufficiently rapidly as for the hydroxyl proton in wet ethyl alcohol (Arnold 1956) or in ammonia (Ogg 1954). The proton exchange between $CH_3NH_3^+$ and H_2O affects the methyl group multiplet structure (Grunwald *et al.* 1956, Ogg and Ray 1957 c, Rocard and Frank 1958). Intermediate spectra change with temperature (Roberts 1956) in a manner formally similar to figure 10. Proton exchange in hydrogen bonded systems has been studied by Schneider and Reeves (1958).

The theoretical line shapes in the presence of nuclear exchange have been calculated (Gutowsky *et al.* 1953, Solomon and Bloembergen 1956).

Internal rotation may modify the spectra considerably, not only by altering chemical shifts, as discussed in § 2.3.5, but by modifying the J values (Shoolery and Crawford 1957). If the occupation times of different orientations are different some care is required in the averaging of J values (Graham and Waugh 1957).

3.5. *Double Resonance in High Resolution Spectra*

A valuable technique which has both a physical interest and great value in facilitating the interpretation of complex spectra is to irradiate one set of equivalent nuclei at its own resonant frequency while observing another set. As already stated, any agency which causes sufficiently rapid transitions between the states of a nucleus will average out its J-type interaction with another nucleus. Rapid transitions between states means that in equation (31), for example, the average value of M_A (or M_B) must be used and this is zero. There will therefore be a collapse of the multiplet splitting of the observed nucleus which was due to the irradiated one. The irradiation of the second nucleus is much more intense than the radiation applied to the nucleus under observation. This is often loosely termed 'saturating' the second nuclear system, but although saturation, i.e. almost equalization of energy level populations, will often occur, it is not the essential requirement. (There may be effects due to the population changes, the Overhauser (1953) effect, for example, which we shall not discuss in this review.) The effect of a large rotating radio-frequency field of strength H_{1B} at the resonant frequency on a nuclear magnetization (of gyromagnetic ratio γ_B) is to nutate it at the rate $\gamma_B H_{1B}/2\pi$ c/s so that the condition for collapse of a multiplet of A due to B is simply $\gamma_B H_{1B}/2\pi \gg J_{AB}$.

Alternatively, consider a coordinate system rotating at ν_B c/s, the resonant frequency of nuclei B. The effective field at the B nuclei is then simply H_{1B} apart from an effective field due to the A nuclei, from the term $-hJ_{AB}\mathbf{I}_A.\mathbf{I}_B$ in the Hamiltonian, which is of order of magnitude $2\pi J_{AB}/\gamma_B$. Thus if H_{1B} is much greater than this field, i.e. if $\gamma_B H_{1B}/2\pi \gg J_{AB}$, the magnetization of the B nuclei is directed along H_{1B}, i.e. in the xy plane. However, the magnetization of the A nuclei which are only weakly irradiated by H_{1A} and are unaffected by H_{1B} are still subject to a static field $H_0 - 2\pi\nu_B/\gamma_A$ which is directed along the z axis. Hence $\langle \mathbf{I}_A \rangle$ is perpendicular to $\langle \mathbf{I}_B \rangle$ and the term $-hJ_{AB}\mathbf{I}_A.\mathbf{I}_B$ contributes nothing to the Hamiltonian and the multiplet is collapsed. An analysis based on the Bloch equations is given by Anderson (1956a). The method was suggested in magnetic resonance by Bloch (1954) and exploited by Bloom and Shoolery (1955); see also Anderson (1956a).

The effect of irradiation at the ^{31}P frequency on the ^{19}F doublet in PO_3F^- is shown in figure 24. An intermediate state between doublet and singlet occurs when $H_{1^{31}P} \simeq \frac{1}{4}$ gauss, i.e. $\gamma_{^{31}P} H_{1^{31}P}/2\pi \simeq 430$ c/s, which is comparable with $J_{^{19}F\text{-}^{31}P} = 860$ c/s in this ion.

This method can be used to determine the source of broadening or multiplicity of a resonance line and gives not only the nucleus to which the observed one is coupled, as identified by the irradiation frequency used, but also an idea of the J value of the coupling via both the change in the observed signal and the H_1 intensity required to produce the change. An absorption line may be broadened by more than one other nuclear species, and then the effect of each can be revealed separately, as shown in figure 25. The proton resonance in aluminium borohydride $Al(BH_4)_3$ is a broad line without structure (figure 25, A). Irradiation at the ^{11}B frequency considerably reduces the width and reveals some indication of structure (figure 25, B) showing the presence of an indirect coupling with the (81%) abundant ^{11}B nuclei. Irradiation at the ^{27}Al (100% abundant) resonant frequency (figure 25, C) shows a more spectacular effect, leaving principally a

1,1,1,1 spectrum which is expected from the ^{11}B nucleus ($I = 3/2$) interacting with the protons ($J_{^1H\,^{11}B} = 87$ c/s). The less intense lines falling almost exactly between the main ones are due to ^{10}B ($I = 3$). These should have a J value differing by the factor $\gamma_{^{10}B}/\gamma_{^{11}B} = 0{\cdot}33_6$ and a 1,(1),(1),1,(1),(1),1 spectrum of intensity $(19/81) \times (4/7) = 0{\cdot}13_4$ of that for interaction with ^{11}B. The lines in parentheses are swamped by the ^{11}B lines. The ^{11}B resonance (figure 25, D)

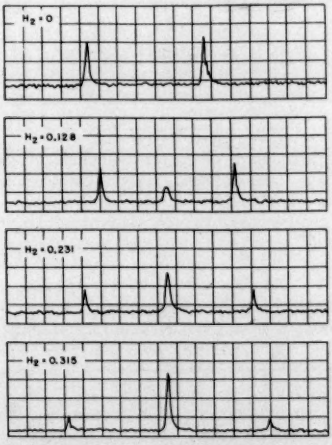


Figure 24. Collapse of the ^{19}F doublet in PO_3F^- on irradiation of ^{31}P (*after* Bloom and Shoolery 1955).
Note: H_3 corresponds to $H_{^{11}B}$ in text.

shows a five-line spectrum 1,4,6,4,1 with spacing 87 c/s to be expected for interaction with four equivalent protons, evidently within a BH_4 group with the protons in a tetrahedral configuration. The interaction of the boron with the aluminium is either small or reduced by an exchange or other mechanism (Ogg and Ray 1955), very probably quadrupole relaxation since ^{27}Al has a large electric quadrupole moment ($0{\cdot}149 \times 10^{-27}$ cm^2). The value of the double resonance technique is further illustrated by a study of formamide ($HC{=}ONH_2$) and its deuterated derivatives (Piette *et al.* 1958).

Double resonance methods have been less exploited for investigating interactions between identical but non-equivalent nuclei. This is presumably because of the experimental difficulty of irradiating strongly at a frequency comparatively close to the one at which observations are made. Anderson (1956 b) has performed this experiment for the protons in dichloroacetaldehyde ($CHCl_2CHO$), obtaining a result completely analogous to figure 24 except that the frequency difference between the two irradiating fields was varied for successive sweeps through resonance. This incidentally is a rather accurate way of measuring the

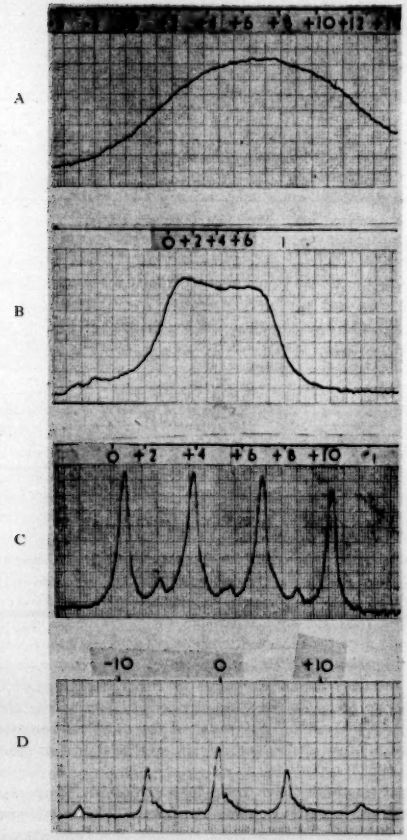


Figure 25. Spectra of aluminium borohydride ($Al(BH_4)_3$) (Ogg and Ray 1955). A, proton resonance at 30 Mc/s; B, as A but ^{11}B irradiated at 9·63 Mc/s; C, as A but ^{27}Al irradiated at 7·82 Mc/s; D, ^{11}B spectrum at 12·3 Mc/s. Scales indicated are in parts in 10^6 at the frequency of observation.

multiplet splitting. Adequate irradiation may not be permissible since we must have $\gamma_B H_{1B}/2\pi \ll \delta_{AB}$ if only one set of equivalent nuclei is to be irradiated. The method would appear, therefore, to be applicable only if $J_{AB} \ll \delta_{AB}$. This, however, is not a serious restriction since if J_{AB} is comparable with δ_{AB} one does not have two separate groups of nuclei to irradiate.

Line shapes under conditions of double resonance have been studied by Rempel (1958) (see also Bloch 1958).

The double resonance method may provide an accurate measure of the magnetic moment of a nucleus for which a direct measure is difficult. In the metal alkyls, for example in lead tetraethyl ($Pb(C_2H_5)_4$), Baker (1957) found that the methyl group resonance is split by interaction with ^{207}Pb. (We note, in passing, that the indirect coupling of ^{207}Pb with the intervening methylene protons is less than with the methyl ones.) Irradiation of ^{207}Pb collapses the methyl signal and gives the ratio $\nu_{^{207}Pb}/\nu_H$ to six figures. This determines the γ value for this only 21% abundant weakly magnetic isotope to a high degree of accuracy in a readily available sample. Most nuclear gyromagnetic ratios have now been determined rather less accurately in one way or another. An extension of this method would seem to provide a readily obtained improvement in accuracy whenever this is demanded by nuclear theory.

3.6. *Analysis of Complex Spectra*

One soon gets quite complex spectra and even when the principal interactions and their order of magnitude may be guessed getting a good fit readily justifies the use of a computer (Heeschen and Gutowsky 1959).

In many cases it is helpful to check the number of distinct interactions for a given molecule and Wilson (1957) has used group theoretical methods. Further simplification is possible by judicious use of the symmetry properties of the nuclear spin wave functions (McConnell *et al.* 1955) and of 'constants of motion' in addition to the usual ones. For example, for a group of equivalent nuclei the square of the total spin angular momentum of the group and its z component are good quantum numbers (Hahn and Maxwell 1952, McConnell *et al.* 1955). We have already noted that one may be considerably helped by spectra obtained at different magnetic fields (figure 20). At high fields second-order splitting is reduced. At low fields some chemical shifts may be effectively zero, thus simplifying the spectra (e.g. Roux 1958). δ is proportional to H_0 and J is independent of H_0. In many cases groups of nuclei only interact weakly if at all, thus giving spectra which are merely superpositions. Double resonance experiments may reveal the origin of a line broadening or multiplicity (Bloom and Shoolery 1955, Anderson 1956 a, Bloch 1956).

A powerful method of analysis is an extension of that widely used for broad magnetic resonance lines. Van Vleck (1948) has shown that the considerable difficulty in obtaining the exact spectrum in complex cases can be partially circumvented by noting that the set of moments of the absorption line $\langle(\nu-\nu_0)\rangle$, $\langle(\nu-\nu_0)^2\rangle$, etc., equally well describe the line. Experimentally, however, it is rarely possible to determine moments higher than the fourth with any confidence because of the large contribution of the 'tails', which tend to be weak signals readily lost in the noise. Further, many lines are symmetric about a centre, so that the odd moments are zero.

An attractive feature of the moment method in the present instance is that it will be recalled that the second moment is independent of 'exchange' terms in the Hamiltonian, i.e. of terms of the form $A_{12}\mathbf{I}_1.\mathbf{I}_2$. Although the J coupling is not physically an exchange term, except in so far as an exchange interaction between the intermediary electrons may be involved, it has this form, and so the second moment depends only on the chemical shifts. The fourth moment does depend on exchange terms (e.g. exchange narrowing and broadening in electron resonance spectra). These statements are illustrated by our simple case of two non-equivalent nuclei with indirect coupling for which the exact spectrum is given in equation (34). Anderson and McConnell (1957) have adapted Van Vleck's theory to the present situation. The part of a spectrum due to a group of nuclei may be separated out if all the lines in the group arise from magnetic transitions within it. The greater the division into groups which is possible, the simpler the spectrum and the more powerful is the moment method. Nuclei of different species always fall in separate groups. If there are only two inequivalent sets of nuclei in the group numbering N_1 and N_2, the chemical shift ${}_{\mathrm{A}}\delta_{12}$ between (1) and (2) is given by

$$\delta_{12}{}^2 = \frac{(N_1+N_2)^2}{N_1 N_2}\langle(\delta\nu)^2\rangle, \qquad \ldots\ldots(37)$$

where $\delta\nu$ is measured from the centre of the line.

This formula is readily confirmed for the examples we have studied and in particular for equation (34). The independence of the second moment on J is consistent, for instance, with the observation that when $J \ll \delta$ the changes in intensity are of second order while the shifts are of first order in J/δ. A valuable check on the accuracy of spectra calculated in detail by the direct method is to calculate the second moment and confirm that it is independent of any J coupling.

Again, for a single J coupling constant we have, when $N_1 = N_2 = 1$ and $I_1 = I_2 = \frac{1}{2}$,

$$J^2 = [\langle(\delta\nu)^4\rangle - \langle(\delta\nu)^2\rangle^2]/\langle(\delta\nu)^2\rangle, \qquad \ldots\ldots(38)$$

and it is readily shown that this expression is correct for equation (34).

Anderson and McConnell (1957) analyse a more complex example, the proton signal of o-dichlorobenzene, shown in figure 26, which is an A_2B_2 spectrum in which the A's and B's are not equivalent. We may assume interactions with the ^{35}Cl nuclei removed by quadrupole relaxation. By symmetry,

$$J_{a1b1} = J_{a2b2} \equiv J_\alpha \quad \text{and} \quad J_{a1b2} = J_{a2b1} \equiv J_\beta.$$

$J_{a1a2} \equiv J_a$ is probably much smaller than $J_{b1b2} \equiv J_b$ and indeed than J_α. The spectrum of figure 26 may be understood by means of the sequence of spectra shown in figure 27 in which suitable relations between the parameters are gradually developed. Pople *et al.* (1957) give an exact analysis of this type of spectrum.

Application of equation (37) to figure 26 gives $\delta_{ab} = 7{\cdot}8 \pm 0{\cdot}5$ c/s, whereas the two apparently most important lines are only 2·2 c/s apart, thus indicating that J and δ are similar in magnitude. The fourth moment is in fact independent of J_a and J_b, and a formula analogous to (38) gives $(J_\alpha{}^2+J_\beta{}^2)^{1/2} = 8{\cdot}1 \pm 1{\cdot}0$ c/s. In this case no further information can be obtained by the moment method unless one goes to higher moments, which is impracticable experimentally and for which the exact theory is not appreciably more difficult. The moment method therefore

yields δ_{ab} and also a good value of J_a since most probably $J_\beta \ll J_\alpha$. A more detailed calculation may be expected to yield J_b on the assumption $J_\beta = J_a = 0$, and improvement of the fit should give finally J_β and J_a if the spectrum is well enough resolved. Hence all the parameters are determined apart from some uncertainty

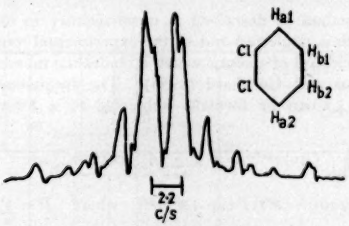


Figure 26. Proton spectrum of o-dichlorobenzene ($C_6H_4Cl_2$) at 30 Mc/s (Anderson and McConnell 1957).

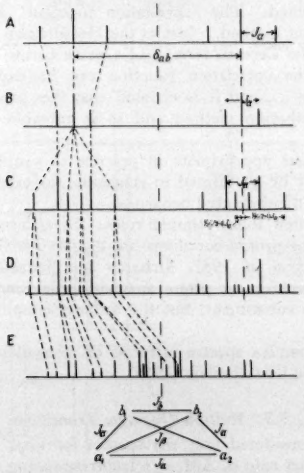


Figure 27. Spectra for the system A_2B_2 for various special values of the parameters, shown qualitatively (based on Pople *et al.* 1957).

A. $J_a = J_\beta \ll \delta_{ab}$, any J_a and J_b. Simple 1, 2, 1 spectrum for two pairs of equivalent nuclei.

B. $J_\beta = J_a = 0$, $J_a \ll J_b \ll \delta_{ab}$. Note same spectrum as A but spacing has different meaning.

C. $J_\beta = J_a = 0$, $J_a = J_b \ll \delta_{ab}$.

D. $J_\beta = J_a = 0$, $J_a \simeq J_b \ll \delta_{ab}$.

E. $J_a \simeq J_b \simeq \delta_{ab}$ and $J_\beta \ll J_a$, $J_a \ll J_b$. This spectrum is for conditions as found in o-dichlorobenzene whose spectrum is shown in figure 26.

in relative signs and complete uncertainty as to the absolute signs of both the J's and δ_{ab} in this example.

The validity and utility of the decomposition of a spectrum into sub-groups is discussed in some detail with a number of examples by Anderson and McConnell (1957).

The moment method as described is unsatisfactory in that the natural and apparative line width is neglected and so the experimental moments are too high.

A more general method of analysis which includes the moment method has been suggested by Primas and Günthard (1958). The frequency spectrum written $Q(\nu) = \Sigma A_{mn}\,\delta(\nu - \nu_{mn})$ can be formally subjected to a Fourier transformation, giving

$$K(t) = 2\pi \int_{-\infty}^{\infty} Q(\nu)\, e^{2\pi i \nu t}\, d\nu = 2 \sum_{m<n} A_{mn} \cos 2\pi \nu_{mn}\, t,$$

$$= \mathrm{Tr}[\exp(-i\mathscr{H}t)\, P \exp(i\mathscr{H}t) P] \quad \text{where} \quad P = \sum_k \gamma_k I_{k,\,x \text{ or } y}.$$

It is readily shown that $K(t)$ is independent of a term in the Hamiltonian for J coupling between equivalent nuclei and hence so is the spectrum, a result which has already been quoted. The 'correlation function' $K(t)$ is the generating function of the moment method. Just as the Hamiltonian may often be regarded as a main part (e.g. the Zeeman term) and a perturbation term (e.g. J coupling), $\mathscr{H} = \mathscr{H}_m + \epsilon\mathscr{H}_J$, so the correlation function can be expanded in a series in $K(t) = K_0(t) + \epsilon K_1(t) + \ldots$, and it is claimed that this procedure is less onerous than the usual perturbation method and so is valuable for the calculation of complex spectra.

Many complex cases approximate in practice to simpler situations and here the interpretation may be performed in stages as, for example, by Richards and Schaefer (1958 b) for trisubstituted benzenes.

Actual J values often follow simple rules. In asymmetrically trisubstituted benzene derivatives the proton couplings are $J_{\text{ortho}} = 7$ to 9 c/s, $J_{\text{meta}} \simeq 2{\cdot}5$ c/s and $J_{\text{para}} < 1$ c/s (Gutowsky *et al.* 1957, Richards and Schaefer 1958 b). Again, in para disubstituted benzenes, the proton magnetic shielding (chemical shift) varies considerably with the substituent, but the indirect coupling does not (Richards and Schaefer 1958 a).

The analysis of complex spectra in terms of molecular structure is discussed by Shoolery (1955) and Chamberlain (1959).

3.7. *Multiple Quantum Transitions*

We have so far considered only the spectra for weak excitation ($H_1 \to 0$), in which case the selection rule is, $\Delta M_z = \pm 1$ corresponding to absorption or emission of a quantum of circularly polarized radiation of angular momentum $\hbar$. When H_1 is finite otherwise 'forbidden' transitions may occur, for example two quanta may be absorbed requiring $\Delta M_z = +2$ for angular momentum balance. Similarly, an n quantum transition may occur if the system has energy levels such that a transition with $\Delta M_z = \pm n$ is possible. Multiple quantum transitions in high-resolution spectra are of particular interest when the single quantum line has structure as we now show.

We need to consider at least two nuclei with $I = \frac{1}{2}$ in order to have $\Delta M_z = \pm 2$. If they do not interact the energy level diagram is shown in figure 28(*a*). The double quantum transition is shown as D, i.e. $2h\nu_D = 2\Delta E$, so that $\nu_D = \nu_s$, the single quantum transition frequency, and so ν_D is not observed separately. This

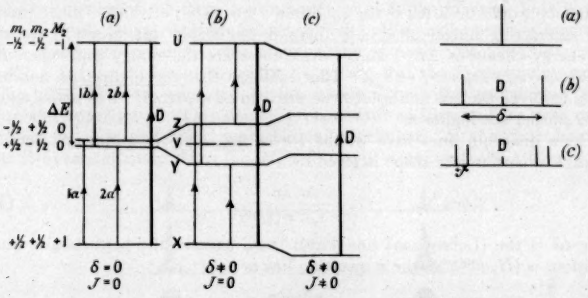


Figure 28. Energy level diagrams and spectra of a two spin (AB) system showing a double quantum (D) transition.

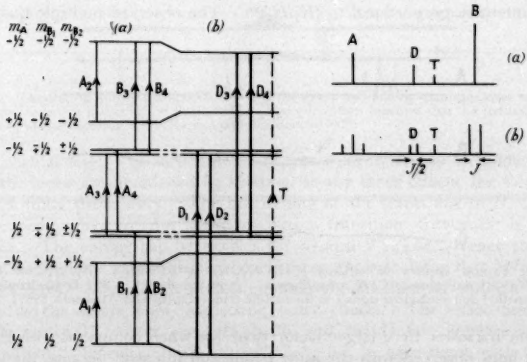


Figure 29. Energy level diagrams and spectra of an AB_2 spin system showing double (D) and triple (T) quantum transitions.

merely reflects the inaccuracy of the previously calculated transition probabilities when H_1 is not small. For two chemically shifted nuclei the double quantum transition occurs for irradiation at $(\nu_1+\nu_2)/2$ (figure 28(*b*)).

With multiplet splitting and the energy levels of figure 28(*c*) only one double quantum transition is possible, giving a single D line at the central frequency.

This double quantum line has been observed in 2-bromo-5-chlorothiophene by Anderson (1956 b). The double quantum line has less structure than the S lines (in this case none). This statement is generally true. For 1,1,2 trichloroethane (AB_2) the energy level diagram is shown in figure 29. The D transitions give a doublet with spacing $J/2$ (assuming $J \ll \delta$), and the D lines are degenerate. This degeneracy would be lifted if the condition $J \ll \delta$ were not valid. Since there are three interacting nuclei of spin $\frac{1}{2}$ there is one triple quantum (T) transition. The energy change is $\Delta E_A + 2\Delta E_B$ where these are the energy gaps when $J = 0$, and so $\nu_T = \nu_A + 2\nu_B = \nu_A + \frac{2}{3}\delta_{AB}$. This confirms that the higher the number of quanta involved the less multiplet structure is to be expected. The actual splitting is also smaller the higher n.

The n quantum lines have widths $1/n$ of the single quantum line because if the single quantum line shape is given by

$$\frac{\Delta\nu/2\pi}{(\nu-\nu_s)^2+(\frac{1}{2}\Delta\nu)^2} \qquad \ldots\ldots(39)$$

where $\Delta\nu$ is the (Lorentzian) line width then, because the perturbation causing transitions is $(H_1 e^{-2\pi i\nu t})^n$, the n quantum line is

$$\frac{n\Delta\nu/2\pi}{n^2(\nu-\nu_n)^2+(\frac{1}{2}\Delta\nu)^2}$$

and therefore a line of width $\Delta\nu/n$. Similarly, other factors apart, an n quantum line has intensity proportional to $(H_1/H_0)^{2n}$. The observed multiple quanta lines

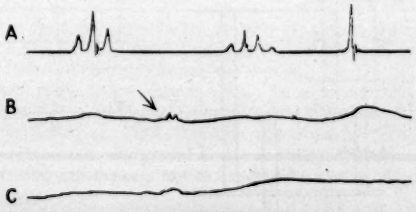


Figure 30. Proton spectra of ethanol (CH_3CH_2OH) at 31·6 Mc/s. A, $H_1 < 0{\cdot}1$ milligauss. B, $H_1 = 3{\cdot}4$ milligauss. C, $H_1 = 6$ milligauss. Note the sharp 1, 4, 4, 1 D quadruplet between the methyl and methylene signals in the middle trace (Kaplan and Meiboom 1957).

often seem narrower by a larger factor than $1/n$ when compared with the single quantum lines observed with the same irradiation intensity because the latter are lifetime broadened by the intensity of H_1 required to make the multiple quantum transitions sufficiently probable.

Multiple quantum transitions are better known in atomic and molecular beam investigations (Grabner and Hughes 1950, 1951, Brossel, Cagnac and Kastler 1954, Kusch 1954). Their possibility in nuclear magnetic resonance was suggested by Hughes and Geiger (1955) and they were observed by Anderson (1956 b). A particularly convincing case in a liquid is given by Kaplan and Meiboom (1957) using the classic material 'wet' ethyl alcohol, which clearly shows the effects we

have discussed. Figure 30, trace B, shows a sharp 1,4,4,1, quadruplet between the broadened methyl and methylene proton signals. The rules for D multiplets are not the same as for S multiplets since even when $J \ll \delta$ the states of at least two groups of non-equivalent nuclei are involved. The authors see triple quantum lines in trace C at the 1/3 and 2/3 positions between the CH_3 and CH_2 signals.

The calculation of the form and intensity of multiple quantum lines is given by Besset *et al.* (1954) and by Salwen (1955). We shall obtain a qualitative result. Ignoring population considerations, the transition probabilities are the controlling factor. Referring to figure 28, we see that the D absorption process may be regarded as the excitation of the system first to the virtual level V and then to the upper level. This double process with an intermediate state can be described by a second-order perturbation calculation. If the four levels in figure 28(*b*) are infinitely sharp (i.e. $H_1 = 0$) the probability of the system getting

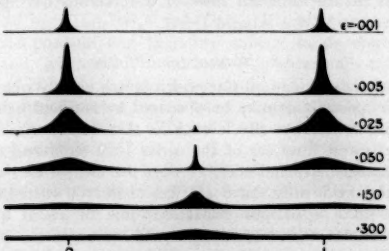


Figure 31. Theoretical absorption spectra showing the single and double quantum lines as a function of the radiation intensity $\epsilon = \frac{1}{2}\gamma H_1/2\pi 2\delta$. The calculation assumes that the natural line width is 1% of the chemical shift (Kaplan and Meiboom 1957).

into the virtual level V is zero and the D transition is strictly forbidden. However, if the levels are broadened by lifetime, or any other effects, the D transition acquires a finite probability. The broadening of the levels due to H_1 the actual radio-frequency field at the single quantum transition frequency is of order $\frac{1}{2}hH_1\gamma/2\pi$. The energy gap between Z (or Y) and V is $\frac{1}{2}h\delta$. Hence the level V is almost as available as the original levels when H_1 is so large that $\frac{1}{2}\gamma H_1/2\pi \simeq \frac{1}{2}\delta$, i.e. $\gamma H_1/2\pi\delta \simeq 1$. The double transition lines should then have much the same intensity as the single ones, neglecting other effects. The exact behaviour is shown in figure 31, where $\epsilon = \frac{1}{2}\gamma H_1/2\pi 2\delta$, and the D line appears with comparable intensity when $\epsilon \simeq 0{\cdot}05$, i.e. $\gamma H_1/2\pi\delta \simeq 0{\cdot}2$ in agreement with our order of magnitude estimate. For a more complete calculation using second-order perturbation theory, see Hughes and Geiger (1955).

A complete description of the effect must also include relaxation effects. The line shape and intensity of the S lines are affected by saturation at the high H_1's employed. So are the D lines, since the observed signal depends on the populations of the extreme levels (e.g. X and U in figure 28). These are modified when $\gamma H_1/2\pi \sim \frac{1}{2}\delta$, since then the S levels are irradiated to a considerable extent at the same time. Finally, as H_1 increases, the D line itself becomes broadened

and subject directly to relaxation effects. In such a complex interconnected level system a simple description is hazardous. Figure 31 has been calculated on the assumption that $\tau\delta = 100$ where τ is a relaxation time. Roughly this means that the natural line width of the S transition (for $H_1 \to 0$) is taken as 1% of the chemical shift and $T_1 = T_2$. A calculation employing Bloch's (1956) density matrix method has been used in obtaining figure 31.

The possibility of multiple quantum transitions must be included in the interpretation of saturation and relaxation effects in apparently single quantum spectra, and such effects have been observed. Anderson (1956 b) has shown that the inner lines of the spectrum of 2-bromo-5-chlorothiophene (figure 21) saturate more readily than the outer ones in a manner explained by Bloch's theory (1956) and that for more intense radiation a sharp double quantum line appears at the centre. This is the D line of figure 28(*c*). In particular, it could lead, together with other causes, to the different lines of a spectrum having different widths (Bloch 1956, Anderson 1956 a, Arnold 1956).

3.8. *Natural Line Widths*

It seems likely that the natural line width is not usually observed even in the highest resolution spectra. It may be obscured by residual field inhomogeneity, H_1 broadening and extraneous effects such as that due to dissolved atmospheric oxygen. The narrowest lines are of the order $\frac{1}{2}$ c/s wide and even for this one requires a field homogeneity and stability over the sample of 1 in 10^8 at 40 Mc/s and the irradiating field must have H_1 less than 100 microgauss for protons. $\frac{1}{2}$ c/s corresponds to a spin–spin relaxation time of about $\frac{1}{3}$ second, whereas in transient experiments relaxation times up to tens of seconds are observed in pure mobile liquids and so the true line widths are very much smaller (see § 4). The line width may differ for different component lines in a high resolution spectrum. An example of this may be seen in figure 19, where the central doublet of the methyl proton signal has one line considerably sharper than the others, although the integrated intensities are the same. These lines correspond to the central ones of figure 18, although that is a simpler case. The sharper line corresponds to the singlet state of the methylene protons with $m_z = 0$, $\mathbf{m} = 0$, whereas the broader one is for the triplet state $m_z = 0$, $\mathbf{m} = 1$. Since transitions out of the singlet state are much rarer than from one triplet state to another, the singlet line is subject to much less lifetime broadening.

In the steady state high resolution experiment one has the interesting possibility of saturating individual components of a complex line and observing the recovery of the signal under a reduced irradiation intensity which is insufficient to saturate. The recovery is at the rate of the spin–lattice relaxation time and the true line width may be deduced from this. This experiment has been performed on toluene ($C_6H_5CH_3$) by Nederbragt and Reilly (1956) for the comparatively well separated signals from the methyl and aromatic protons, and it is found that $T_{1CH_3} = 9$ sec and $T_{1C_6H_5} = 16$ sec at 25°c corresponding to line widths of about 10^{-2} c/s. These results are confirmed by an apparent T_1 of 12·5 sec observed by a transient method (Giulotto 1956). The interpretation of such results is complicated by the fact that the two sets of protons are not entirely independent systems although the coupling in this case must be rather weak.

These observations may be more significant for the theory of liquids than that of molecules, as may be seen from the effect of dilution in a solvent which does not contain magnetic nuclei. For example, in p–xylene ($(CH_3)_2C_6H_4$) the ring and methyl proton spin–lattice relaxation times are 14 and 7·5 seconds respectively. In 11% solution in carbon bisulphide they are 37 and 10 seconds, showing the effect of intermolecular interactions. A similar effect is found in other materials (Nederbragt and Reilly 1956), and is particularly striking in mesitylene ($(CH_3)_3C_6H_3$) in CS_2 (Reilly and Strombotne 1957). The relaxation in dilute solution is satisfactorily explained by direct dipole interaction within the molecule.

The saturation properties of spectra with a number of components (Portis 1956) which are not independent is complex and we shall not discuss it here since we are not primarily concerned with relaxation (see also Wertz *et al.* 1956).

Lifetime broadening due to quadrupole relaxation is usually the factor controlling the line width when the nucleus has spin greater than $\frac{1}{2}$. The ^{14}N resonance is broad in $(CH_3)_2NH_2^+$ and $(CH_3)_3NH^+$ where the nitrogen nucleus is in an asymmetric position and therefore subject to an electric field gradient. On the other hand, when the ^{14}N nucleus is in a centrally symmetric situation, as in NH_4^+ and $(CH_3)_4N^+$, it shows a narrow line (Ogg and Ray 1957 a). Again, the comparatively broad proton resonance in ammonia $^{14}NH_3$ is due to quadrupole relaxation of ^{14}N since in $^{15}NH_3$ (^{15}N has $I = \frac{1}{2}$ and therefore $Q = 0$) the proton line is sharp (Ogg and Ray 1957 b).

The approach to higher and higher resolution may be frustrated by radiation damping (Bloembergen and Pound 1954, Bruce, Norberg and Pake 1956). The observation of a resonance requires essentially that a transverse magnetization induces a voltage in a coil. The nuclear system must therefore provide the energy for the signal and so nuclear transitions occur. Any resonance line is therefore subject to lifetime broadening of a radiating system. Because of low filling factors and numerous lines in high-resolution spectra, the radiation damping line width is probably much smaller than the resolution at present achieved. This effect is also present in transient experiments but is minimized by the echo method described in § 4.3 (Powles and Cutler 1957, 1958).

§ 4. Magnetic Shielding and Indirect Spin–Spin Interaction Effects in Transient Experiments

4.1. *Introduction*

We have noted the serious experimental difficulties in the steady state experiments in producing a sufficiently homogeneous field. One way round this difficulty is to do a transient experiment. The philosophy is rather similar to the two ways of investigating a passive electrical circuit, that is, either by measuring amplitude and phase relationships on the application of sinusoidal signals, or by applying a short pulse and observing the transient signals as a function of time. The transient method has both advantages and disadvantages. All the effects discussed in previous sections have their counterpart with some others besides.

4.2. *Decay Signals*

The simplest transient experiment is to produce a nuclear magnetization in the plane transverse to the main field direction Oz. This magnetization, in the

absence of a driving field H_1, precesses about the Oz axis with its resonant frequency $\nu_r = \gamma H_0/2\pi$ and induces an alternating voltage at this frequency in a coil placed with its axis at right angles to H_0, and this signal is observed. The transverse magnetization may be produced by the application of a pulse of radio-frequency field at the resonant frequency lasting for a time $t_\omega \ll T_1, T_2, T_2^*$ where T_2^* is the relaxation time corresponding to the line width produced by field inhomogeneity (equation (6)). The applied field nutates the magnetization by the angle $\gamma H_1 t_\omega$ radians and produces transverse magnetization of magnitude M_0 if $\gamma H_1 t_\omega = \pi/2$, i.e. a 90° pulse. In some cases it is convenient to polarize the nuclei with a comparatively large field (several hundred gauss) and remove it, leaving the polarization in a field (often the earth's field of about one half gauss) at right angles to the original one (Packard and Varian's free precession experiment 1954). A line having the Lorentzian shape (equation (39)) in a steady state experiment, as is usually the case for a liquid, gives after a 90° pulse a signal of frequency ν_r decaying exponentially with time constant T_2, i.e. the Fourier transform signal, and this is in agreement with the appropriate solution of the Bloch equations (1946). If the line shape is not Lorentzian, as in solids and viscous liquids, the considerations are more complex. In order to observe such decays directly the field homogeneity must be such as would allow the observation of the true line width in a steady state experiment.

For more than one set of equivalent nuclei which are not interacting with each other, i.e. $J = 0$, the separate magnetizations may be treated independently. Apply a 90° pulse to both sets of nuclei. This is readily done at radio frequencies when the nuclei are identical but non-equivalent and less easily accomplished if they are different nuclei (Hahn and Kaplan 1958). At low frequencies the situation is more favourable (Elliott and Shumacher 1957, Powles and Cutler 1958). The two, or more, magnetizations precess at their own Larmor frequency, decaying at a rate determined by the individual line width, and so a coil and detector system sensitive to both signals will observe a decaying beat signal, or, in other words, a signal at an intermediate frequency which is amplitude modulated at the difference (chemical shift) frequency

$$M_{xy} = M_{01} \exp\left[2\pi i(\nu_m - \tfrac{1}{2}\delta) - \frac{1}{T_{21}}\right] t + M_{02} \exp\left[2\pi i(\nu_m + \tfrac{1}{2}\delta) - \frac{1}{T_{22}}\right] t.$$

More or less complicated beat patterns will be observed, depending on the number of signals and their relative amplitudes and decay times. Such signals have not been published, since when it is possible to observe them it is possible to do the steady state experiment which is more readily interpreted.

If the inequivalent nuclei have a J-type coupling, the Bloch equations are inadequate to describe the situation, and we revert to the quantum mechanical calculation.

The time dependent wave function for $I = \frac{1}{2}$ is

$$\psi = a\alpha \exp(2\pi i \nu_r t/2) + b\beta \exp(-2\pi i \nu_r t/2)$$

and a, b are time independent between pulses (apart from relaxation effects) and time dependent during them. It is sufficient to consider the transverse magnetization which is proportional to $\langle I_+ \rangle$. Thus $V_+ \propto a^* b \exp(-2\pi i \nu_r t)$ is the

voltage induced in transverse coils. Initially the phases of a and b are incoherent and so $V_+(0) = 0$. A magnetization in the Oz direction is represented by taking $a(0) = 1$, $b(0) = 0$, since it is only necessary to consider population differences. The effect of the applied radio-frequency pulse is found by solving the time dependent Schrödinger equation $ih\partial\psi/\partial t = \mathscr{H}\psi$, where perturbing terms in $\mathscr{H}$ other than that due to the pulse may be neglected under the conditions already mentioned. We find

$$a(t) = a(0)\cos(2\pi\nu_1 t/2) + ib(0)\sin(2\pi\nu_1 t/2),$$
$$b(t) = b(0)\cos(2\pi\nu_1 t/2) + ia(0)\sin(2\pi\nu_1 t/2),$$

i.e. a 90° pulse applied to the equilibrium magnetization gives $a(t_\omega) = 1/\sqrt{2}$ and $b(t_\omega) = i/\sqrt{2}$, and so after the pulse, i.e. $t > t_\omega$, $V_+ \propto \frac{1}{2}i\exp(-2\pi i\nu_r t)$ which corresponds to a transverse magnetization rotating at ν_r. The longitudinal magnetization proportional to $(aa^* - bb^*)$ is naturally zero. Decay processes due to incoherent dephasing of transverse magnetization are usually included by arbitrarily introducing a factor $\exp(-t/T_2)$. The signals from inequivalent non-interacting nuclei can be added vectorially.

The simplest system with indirect coupling is that of two single non-equivalent nuclei of spin $\frac{1}{2}$ whose steady state spectrum is given by equation (34) and whose wave functions are given in equation (32) apart from the frequency factors which are readily obtained from the energy levels in equation (33). The Bloch equations are not now applicable since the two magnetizations 'interact' as indicated by the mixing of states. The equilibrium condition, a Boltzmann distribution over the levels, may be taken as $a_{1,-1} = 0$, $a_{0,0} = 1$, $a_{1,0} = e^{i\alpha}$, $a_{1,1} = \sqrt{2}\,e^{i\beta}$, where the time dependent wave function has been written

$$\psi = a_{11}\alpha\alpha\exp\left(\frac{-iE_{11}t}{\hbar}\right) + a_{1-1}\beta\beta\exp\left(\frac{-iE_{1-1}t}{\hbar}\right) + \frac{1}{\sqrt{2}}\left(\frac{a_{00} - a_{10}R}{Q-R}\right)$$
$$\times[\alpha\beta(1+Q) + \beta\alpha(1-Q)]\exp\left(\frac{-iE_0 t}{\hbar}\right) - \frac{1}{\sqrt{2}}\left(\frac{a_{00} - a_{10}Q}{Q-R}\right)$$
$$\times[\alpha\beta(1+R) + \beta\alpha(1-R)]\exp\left(\frac{-iE_0' t}{\hbar}\right).$$

Any observable must be averaged over α and β since the phases of the nuclei are initially random. After a 90° pulse

$$a_{1-1} = -\frac{1}{\sqrt{2}}(e^{i\beta} - ie^{i\alpha}), \quad a_{00} = 1, \quad a_{10} = ie^{i\beta}, \quad a_{11} = \frac{1}{\sqrt{2}}(e^{i\beta} + ie^{i\alpha}).$$

These coefficients can be shown to correspond (Powles, unpublished calculations) to a transverse rotating magnetization at the mean frequency ν_m of amplitude

$$\left|\left[1 + \frac{J}{(\delta^2 + J^2)^{1/2}}\right]\cos\pi[(J^2 + \delta^2)^{1/2} - J]t + \left[1 - \frac{J}{(\delta^2 + J^2)^{1/2}}\right]\cos\pi[(J^2 + \delta^2)^{1/2} + J]t\right|.$$

Such a signal does not seem to have been reported, presumably because if it could be observed the simpler steady state signal could equally readily be obtained.

When $J \ll \delta$ this becomes a signal of amplitude $|2\cos\pi Jt\cos\pi\delta t|$ observing both nuclei.

But when $J \ll \delta$ one can observe the signal from each nucleus separately. These signals are readily shown to be at frequency $\nu_m - \frac{1}{2}\delta$ and $\nu_m + \frac{1}{2}\delta$ each with amplitude $|\cos \pi Jt|$. Hence when $J \ll \delta$ there are in effect four independent transverse magnetizations of equal amplitude with frequencies $\nu_m \pm \frac{1}{2}\delta \pm \frac{1}{2}J$ which may be regarded as signals from four sets of nuclei: one set is nuclei of type (1) with nuclei (2) in the α state, the second set is nuclei type (1) with nuclei (2) in the

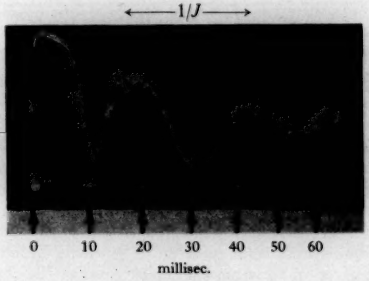


Figure 32. Deuteron decay signal in hydrogen deuteride. (Rectified signal, base-line is indicated by close dots and time from 90° pulse in milliseconds by arrows). (Carr and Purcell 1952).

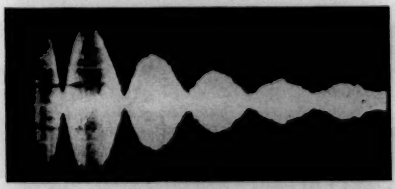


Figure 33. Free precession (2 kc/s) proton signal in fluorobenzene (C_6H_5F) in the earth's field (0·5 gauss). The modulation corresponds to a J value of 5·8 c/s. Total sweep is 1 second (Elliott and Schumacher 1957).

β state and similarly for nuclei (2). Generalizations can readily be made for more numerous types of nuclei and for I greater than $\frac{1}{2}$. We note that the amplitude modulation is not the usual radio one—there is no carrier. A decay factor $\exp(-t/T_2)$ must be added to allow for incoherent dephasing.

When the nucleus causing the indirect coupling is of a different species we certainly have $J \ll \delta$ and one would normally observe only the magnetization due to one nuclear species, and we should see an amplitude modulation $|\cos \pi Jt|$. An example of this is the modulated (rectified) deuteron resonance decay signal in HD shown in figure 32 (Carr and Purcell 1952) from which one readily deduces $J_{HD} = 43$ c/s. The proton signal is a beat signal from magnetization at three frequencies, $\nu_m - J$, ν_m and $\nu_m + J$ of equal amplitude corresponding to the three equally probable states of the deuteron, i.e. the modulation is $|1 + 2\cos 2\pi Jt|$.

Another example is shown in figure 33 (Elliott and Shumacher 1957) for the proton signal in fluorobenzene (C_6H_5F) in the earth's field. The modulation of the form $\cos \pi Jt$ corresponds to an apparent coupling J_{HF} of $5{\cdot}8 \pm 0{\cdot}2$ c/s. The high field steady state proton signal approximates a doublet separated by 6·7 c/s (Baker 1955) so that the free precession signal is the Fourier transform of the steady state signal. The condition $J \ll \delta$ is satisfied since $\delta_{H\,vF}$ even at 2 kc/s is 150 c/s. Higher resolution of the high field spectrum (Williams and Gutowsky 1956, Saika and Gutowsky 1956) resolves the proton signal and shows that the coupling of the ortho, meta and para protons to the fluorine nucleus are 8·5, 6·5 and 2·2 c/s respectively (see also Bak *et al.* 1958). At low fields where $\delta \ll J$ for the protons the steady state spectrum would be a doublet with a separation $\frac{1}{3}(2 \times 8{\cdot}5 + 2 \times 6{\cdot}5 + 2{\cdot}2) = 6{\cdot}4$ c/s and so a modulation frequency of this value would be expected. At $H_0 = 35$ gauss the proton spectrum is a doublet with $J = 6{\cdot}2 \pm 0{\cdot}5$ c/s (Roux 1958). This weighted mean J value is readily understood by taking the simpler case of two protons a and b coupled (J_a and J_b) to one fluorine nucleus, i.e. the perturbing terms in the Hamiltonian are

$$-h[J_a I_{az} I_{Fz} + J_b I_{bz} I_{Fz}]$$

since $J_{a,b} \ll \delta_{HF}$. But since $\delta_{H_aH_b} \ll J_a, J_b$, we must take the zero-order proton wave functions as $\alpha\alpha$, $(\alpha\beta+\beta\alpha)/\sqrt{2}$, $(\alpha\beta-\beta\alpha)/\sqrt{2}$ and $\beta\beta$ with transitions only between the triplet levels. The term in the Hamiltonian splits the extreme levels by $\frac{1}{2}[J_a+J_b]$ and leaves the middle one unaffected, i.e. the proton signal is a doublet with the weighted mean splitting $\frac{1}{2}[J_a+J_b]$. This argument is readily extended to more complicated cases such as fluorobenzene. It would seem that, in view of the difficulty of the steady state experiment, the transient experiment provides a rather more reliable weighted mean value of J, assuming in this case that the indirect coupling is field independent. Very small splittings in the steady state experiment which tend to be obscured by apparative effects appear in the transient experiments as slow modulations. This experiment, in particular, shows the value of experiments at low fields where field homogeneity some orders of magnitude greater than at high fields is possible. The steady state experiment at low fields suffers severely from lack of intensity as well as loss of chemical shift between inequivalent but identical nuclei, and hence of detail.

A rather similar method of measuring indirect coupling due to Reilly (1956) is by using a modification of the 'wiggles' experiment (Bloembergen *et al.* 1948, Jacobsohn and Wangsness 1948). If one passes through resonance rather rapidly ($dH_0/dt > 1/4\gamma T_2^2$) a transverse nuclear magnetization is produced which has an instantaneous frequency which follows the field. This induces a signal which beats with the (constant frequency) driving signal as shown in figure 34, A. This is comparable with the free precession decay signal and so decays as $\exp(-t/T_2)$. However, if the nucleus is indirectly coupled to another nucleus we have noted that the transverse magnetization is modulated with the factor $\cos \pi Jt$ (when $J \ll \delta$). Hence the wiggle amplitude will be modulated by the same factor. (This is for a doublet; if there are n multiplets one readily shows by adding rotating magnetizations that the factor is $\cos^{n-1} \pi Jt$.) This modulation is shown in figure 34, B for the ^{19}F line due to CF_3 ($\cos \pi Jt$), and in C for the ^{19}F line of CF_2 ($\cos^2 \pi Jt$) in CF_3CF_2COOH. Hence $J_{CF_3-CF_2} = 1{\cdot}48$ c/s. This modulation

only appears if the field homogeneity is such that the multiplet would be resolved in a steady state experiment. However, the *stability* requirements are less severe. Note the difference in the modulation due to, say, a 1,1,1 (as for H in HD) and a 1,2,1 (as for CF_3 in CF_3CF_2COOH) multiplet.

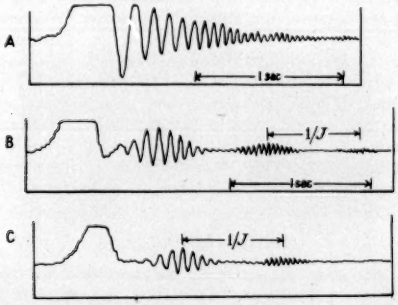


Figure 34. Modulated 'wiggles' after a rapid passage of resonance. A, ^{19}F signal from CF_3COOH, i.e., no indirect coupling, normal wiggles; B, ^{19}F signal from CF_3 group in CF_3CF_2COOH, amplitude modulation $\cos \pi Jt$; C, ^{19}F signal from CF_2 group in CF_3CF_2COOH, amplitude modulations $\cos^2 \pi Jt$ (Reilly 1956).

4.3. *Echoes*

A most significant advance in the transient experiments was Hahn's (1950) method of obtaining echoes. This allows the measurement of true line widths T_2, spin–lattice relaxation time T_1, chemical shifts δ and indirect interactions J, even when the field inhomogeneity is so great as to render these unobservable in a steady state experiment. This constitutes a considerable advance not only because of experimental convenience but because it is desirable to study the relaxing nuclear system in the absence of possible complicating effects produced by the driving field H_1 (Redfield 1955, Tomita 1958). One may also hope for new information in that a transient experiment may provide results which are less averaged over time than the steady state ones. It may be expected, therefore, to distinguish static and dynamic effects in favourable cases (Herzog and Hahn 1956).

The formation of an echo has been explained (Hahn 1950, Pake 1956). In the simplest case one applies a 90° pulse after which there is a signal whose decay is determined by the field inhomogeneity with time constant T_2^*. If a 180° pulse is applied at a time τ later an 'echo' appears centred at 2τ and of width about $2T_2^*$ whose amplitude is proportional to $\exp(-t/T_2)$ which may therefore be used to measure T_2 in spite of the fact that $T_2^* \ll T_2$. Thus the adverse effect of field inhomogeneity is circumvented and indeed it becomes essential to have some present. A system of two non-equivalent independent nuclei with a shift of δ c/s are in no sense different from a rather special inhomogeneity, so that if $1/2\pi T_2^* > \delta$, i.e. if δ is unobservable in a steady state experiment, δ is not observable from the echo. If $1/2\pi T_2^* \simeq \delta$ a given echo signal is observed to have

symmetrical modulation at frequency δ. However, if the sets of nuclei have indirect coupling, the phase of the precessing nuclei between the pulses is modified as described in § 4.2, and so the echo, which depends critically on phase, shows an effect; indeed, in general it shows effects due to both δ and J.

The simplest case is for the 90–180° sequence when it may be shown (Powles, unpublished) that the echo amplitude (at time 2τ) is proportional to

$$2\left|\frac{\delta^2}{(\delta^2+J^2)}\cos\pi J(2\tau)+\frac{J^2}{\delta^2+J^2}\cos\left[\pi(\delta^2+J^2)^{1/2}(2\tau)\right]\cos\left[\pi J(2\tau)\right]\right.$$
$$\left.+\frac{J}{(\delta^2+J^2)^{1/2}}\sin\left[\pi(\delta^2+J^2)^{1/2}(2\tau)\right]\sin\left[\pi J(2\tau)\right]\right|,$$

i.e. the echo amplitude is modulated by both J and δ.

When $J\ll\delta$ this becomes $2\cos\pi J(2\tau)+O(J/\delta)$ and the echo amplitude is modulated by indirect coupling only. If $J=0$ or $\delta=0$ or both are zero, the echo amplitude is not modulated. The result for $J\ll\delta$ may be understood by manipulation of the magnetizations of § 4.2, some care being required for the effect of the pulses. Hence by varying τ, J may be measured. The same result is obtained for the 90°, 180° at τ, 180° at 3τ, etc. sequence (Powles, unpublished).

The echo modulation experiments were initiated by Hahn and Maxwell (1951) and by McNeil *et al.* (1951), who have used the Hahn (1950) 90°, 90° at τ pulse sequence. The echo amplitude envelope for two single inequivalent nuclei observing both is then given (Hahn and Maxwell 1952) by

$$\frac{\delta^2}{2(\delta^2+J^2)}\left|1+\frac{J^2}{2\delta^2}-2\sin^2\pi J\tau\sin^2\pi\tau(J^2+\delta^2)^{1/2}\right|. \qquad \ldots\ldots(40)$$

When $J\ll\delta$, this becomes

$$\left|1-\tfrac{1}{2}[1-\cos\pi J(2\tau)]\,[1-\cos\pi\delta(2\tau)]\right|. \qquad \ldots\ldots(40')$$

Thus no echo amplitude modulation should be observed if either $J=0$ or $\delta=0$, but if both are non-zero both modulation frequencies appear even when $J\ll\delta$, in contrast with the 90–180° sequence.

BrF_5 has two inequivalent sets of ^{19}F nuclei and $J\ll\delta$. The double modulation is shown in figure 35, which is in agreement with equation (40′). When J coupling is present and $1/2\pi T_2^*\simeq\delta$, we can observe modulation of the echo itself (as well as of its amplitude as a function of τ) which has the frequency δ but is now asymmetric about the mean echo envelope (due to field inhomogeneity) (figure 36) because of the J coupling dephasing which causes the two precessing nuclear magnetizations to come into phase not exactly at 2τ except when $\delta.2\tau=n$, i.e. at the turning points of the fast modulation.

Hahn and Maxwell (1952) have studied many cases of echo amplitude modulation in liquids and find that the results are always explicable by terms in the Hamiltonian of the form $-hJ_{12}\mathbf{I}_1.\mathbf{I}_2$, and so are completely in accord with the steady state measurements.

More complex modulation patterns are observed when J and δ are comparable in magnitude. The echo amplitude is also subject to attenuation factors corresponding to the line widths and due to self-diffusion attenuation (Hahn 1950).

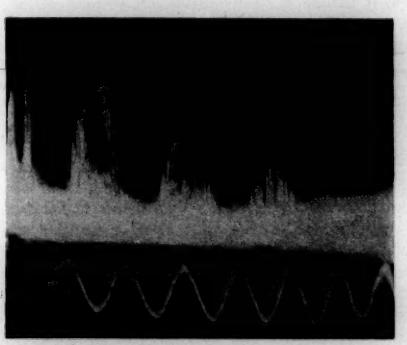


Figure 35. Echo envelope modulation in BrF_5 (BrF_4F) at 30 Mc/s showing two modulation frequencies. One is comparable with the 60 c/s timing wave and is due to J. Each hump suffers a fast modulation δ. Each hump suffers many cycles of δ and each cycle of δ contains many echoes. (Driving pulses are not visible.) (McNeil *et al.* 1951).

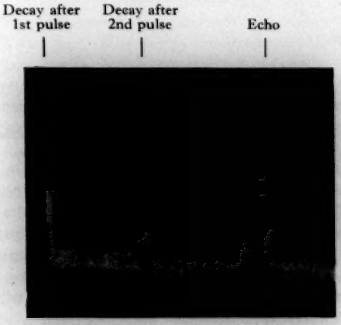


Figure 36. Asymmetrically modulated single echo showing δ and J effects. (McNeil *et al.* 1951).

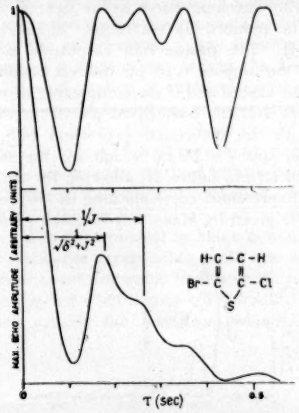


Figure 37. Proton echo *envelope* in 2-bromo-5-chlorothiophene which contains two single non-equivalent protons (AB) and $J \simeq \delta$ at 24 Mc/s. $\delta = 4{\cdot}2$ c/s, $J = 3{\cdot}9$ c/s. Upper curve is the normalized envelope (Hahn and Maxwell 1952).

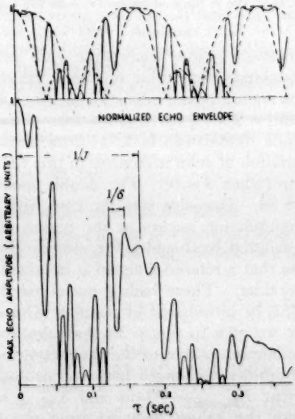


Figure 38. Echo amplitude modulation in 1 : 1 : 2-tri-chloroethane ($CHCl_2CH_2Cl$). AB_2 system for $J \ll \delta$, $J = 6$ c/s, $\delta = 48$ c/s at 24 Mc/s. Note the 2δ modulation (Hahn and Maxwell 1952).

These attenuation effects are usually allowed for in a somewhat arbitrary manner by normalizing the echo envelope curve as in figures 37 and 38. The self-diffusion effect could be reduced by use of the 90, 180, 180, ... pulse sequence (Carr and Purcell 1954). The proton echo amplitude modulation in 2-bromo-5-chlorothiophene is of the simplest type, i.e. indirect coupling between two single protons, although in this case δ and J are comparable in magnitude. Figure 37 shows the echo envelope and the normalized plot from which it will readily be confirmed to agree with the theoretical expression (40) with the parameters $\delta = 4{\cdot}2$ c/s (at 24 Mc/s) and $J = 3{\cdot}9$ c/s which are the same as those deduced from the steady state spectrum, figure 21, allowing for the effect of the different operating field on δ. Expressions corresponding to (40) for more complex cases and further examples are given by Hahn and Maxwell (1952). When J is rather

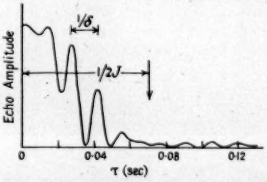


Figure 39. Echo amplitude modulation in wet ethyl alcohol. Arrow indicates point of 'symmetry' at $1/2J$ seconds and is at $\tau = 0{\cdot}071$ second, i.e., $J = 7{\cdot}1$ c/s. Similarly $1/\delta = 0{\cdot}014$ second, i.e., $\delta = 71$ c/s at 29 Mc/s. These are precisely the values of δ and J deduced from the steady state spectrum given in figure 17. Figure 39 contains the same information as the A trace in figure 17, and indeed as in the A trace of figure 19, with the exception of δ_{CH_2OH} (Crawford and Foster 1956).

smaller than δ, one observes a slow modulation with period depending on the J values and a fast modulation depending on the δ values. Some care must be taken since there may be regions of the time scale in which a 2δ modulation appears (one recalls that $|\cos 2\pi\delta t| = 2/\pi + \frac{2}{3}\cos 2\pi . 2\delta t + \ldots$). This effect is seen in figure 38 for the echo amplitude modulation in 1 : 1 : 2-trichloroethane ($CHCl_2CH_2Cl$). It is due to the magnetization at echo maximum in this time region being opposite from the usual direction (when $J = 0$). The double modulation due to J and δ is clearly seen in figure 38. One also sees the ease with which J is determined when $J \ll \delta$. The 2δ modulation occurs in the region of $\tau = (n + \frac{1}{2})/J$ seconds. This effect is readily avoided by loading the system with identical nuclei not subject to J coupling, so that a reference signal is obtained and the magnetization does not 'reverse' at any time. These loading nuclei may be in the same molecule (Hahn and Maxwell 1952) or introduced by solution (Crawford and Foster 1956). As a matter of interest we give in figure 39 the echo amplitude modulation in ethyl alcohol. The modulation is due to the interaction between the methyl and methylene protons, the rapidly exchanged hydroxyl protons merely 'loading' the signal. One deduces that $J_{CH_3-CH_2} = 7{\cdot}1$ c/s and $\delta_{CH_3CH_2}$ at 29 Mc/s is 71 c/s in excellent agreement with the values deduced from the steady state spectrum, figure 17. Figure 39 contains the same information as in figure 17, A, or indeed as in figure 19, A, with the exception of the shift δ_{CH_3-OH}, since J_{CH_3-OH} is nullified

by hydroxyl proton exchange. The attenuation is almost entirely due to self-diffusion and a value for the self-diffusion constant could be deduced from it.

As for the steady state experiment, some interactions which should be present are not observed and for the same reasons. Any interruption of the precession or change in environment of a nucleus may reduce the J and/or δ effects as, for example, by inversion, internal rotation or chemical exchange. The lifetime of a state, or of a nucleus in a particular position, changes effects due to δ or J when it becomes comparable with $1/2\pi\delta$ or $1/2\pi J$. Since usually $J \ll \delta$, the echo modulation tends to disappear, due to loss of the J type interaction. Hahn and Maxwell (1952) have given closer consideration to these effects particularly as applied to the echo experiment and also to the effects of relaxation.

The echo modulation method will be of greatest value when combined with the steady state results. The experimental techniques would benefit from further development. Double resonance experiments could also be performed when the analogue of the 'collapsed' spectrum would be loss of the echo amplitude modulation. The fact that this is not a saturation effect could be neatly illustrated by also pulsing the nucleus causing the J splitting which should not remove the modulation.

References

General References

Andrew, E. R., 1955, *Nuclear Magnetic Resonance* (Cambridge : University Press).
Darrow, K. K., 1953, *Bell Syst. Tech. J.*, **32**, 74.
Grivet, P. (Ed.), 1955, *La résonance paramagnétique nucléaire* (Paris : C.N.R.S.).
Gutowsky, H. S., 1956, *Physical Methods of Chemical Analysis*, Vol. III (New York : Academic Press) p. 303.
Kopfermann, H., 1956, *Kernmomente*, 2nd Edn (Frankfurt : Akad. Verlag). Transl. by Schneider, E. E., 1958 (New York : Academic Press).
Lösche, A., 1957, *Kerninduktion* (E. Berlin : Verb. deutsch Verlag Wiss.).
McConnell, H. M., 1957, *Ann. Rev. Phys. Chem.*, **8**, 105.
Pake, G. E., 1950, *Amer. J. Phys.*, **18**, 438, 473 ; 1956, *Solid State Physics*, Vol. 2, pp. 1–91 (New York : Academic Press).
Powles, J. G., 1956, *Science Progress*, **44**, 449.
Ramsey, N. F., 1953, *Nuclear Moments* (New York : Wiley).
Shoolery, J. N., and Weaver, H. N., 1955, *Ann. Rev. Phys. Chem.*, **6**, 433.
Wertz, J. E., 1955, *Chem. Rev.*, **55**, 829.

Review articles covering similar subject matter to this Review

Abragam, A., 1957, Suppl. to *Nuovo Cim.*, **6**, 1015 (physical aspects).
Gutowsky, H. S., 1958, *Ann. N.Y. Acad. Sci.*, **70**, 786 (chemical aspects).

References

Abragam, A., 1957, *Nuovo. Cim. Suppl.*, **6**, 1015.
Aihara, E., 1957, *J. Chem. Phys.*, **26**, 1347.
Alexander, S., 1958, *J. Chem. Phys.*, **28**, 358.
Anderson, W. A., 1956 a, *Phys. Rev.*, **102**, 151 ; 1956 b, *Ibid.*, **104**, 850.
Anderson, W. A., and McConnell, H. N., 1957, *J. Chem. Phys.*, **26**, 1496.
Arnold, J. T., and Packard, M., 1951, *J. Chem. Phys.*, **19**, 1608.
Arnold, J. T., 1956, *Phys. Rev.*, **102**, 136.
Bader, A. R., Gutowsky, H. S., Williams, G. A., and Yankwich, P. E., 1956, *J. Amer. Chem. Soc.*, **78**, 2385.

Bak, B., Shoolery, J. N., and Williams, E. A., 1958, *J. Mol. Spect.*, **2**, 525.
Baker, E. B., 1955, *J. Chem. Phys.*, **23**, 984 ; 1957, *Ibid.*, **26**, 960.
Banerjee, M. K., Das, T. P., and Saha, A. K., 1954, *Proc. Roy. Soc.* A, **226**, 490.
Becker, E. D., Liddel, U., and Shoolery, J. N., 1958, *J. Mol. Spect.*, **2**, 1.
Bernstein, H. J., Pople, J. A., and Schneider, W. G., 1957, *Canad. J. Chem.*, **35**, 65.
Bernstein, H. J., and Schneider, W. G., 1956, *J. Chem. Phys.*, **24**, 468, 469.
Bernstein, H. J., Schneider, W. G., and Pople, J. A., 1956, *Proc. Roy. Soc.* A, **236**, 515.
Besset, C., Horowitz, J., Messiah, A., and Winter, J., 1954, *J. Phys. Radium*, **15**, 251.
Bloch, F., 1946, *Phys. Rev.*, **70**, 460 ; 1954, *Ibid.*, **93**, 944 (A) ; 1956, *Ibid.*, **102**, 104 ; 1958, *Ibid.*, **111**, 841, see also 853.
Bloembergen, N., and Pound, R. V., 1954, *Phys. Rev.*, **95**, 8.
Bloembergen, N., Purcell, E. M., and Pound, R. V., 1948, *Phys. Rev.*, **73**, 679.
Bloembergen, N., and Rowland, T. J., 1955, *Phys. Rev.*, **97**, 1679.
Bloom, A. L., and Shoolery, J. N., 1955, *Phys. Rev.*, **97**, 1261.
Bothner-By, A. A., and Glick, R. E., 1956, *J. Amer. Chem. Soc.*, **78**, 1071 ; 1957, *J. Chem. Phys.*, **26**, 1647, 1651.
Bottini, A. T., and Roberts, J. D., 1956, *J. Amer. Chem. Soc.*, **78**, 5126.
Brossel, J., Cagnac, B., and Kastler, A., 1954, *J. Phys. Radium*, **15**, 6.
Brownstein, S., 1958, *J. Amer. Chem. Soc.*, **80**, 2300.
Bruce, C. R., Norberg, R. E., and Pake, G. E., 1956, *Phys. Rev.*, **104**, 419.
Carr, H. Y., and Purcell, E. M., 1952, *Phys. Rev.*, **88**, 415 ; 1954, *Ibid.*, **94**, 630.
Chamberlain, N. F., 1959, *Analytical Chemistry*, **31**, 56.
Cohen, A. D., and Reid, C., 1956, *J. Chem. Phys.*, **25**, 790.
Cohen, A. D., Sheppard, N., and Turner, J. J., 1958, *Proc. Chem. Soc.*, 118.
Corey, E. J., Burke, H. J., and Remers, W. A., 1955, *J. Amer. Chem. Soc.*, **77**, 4941.
Corio, P. L., and Dailey, B. P., 1956 a, *J. Amer. Chem. Soc.*, **78**, 3043 ; 1956 b, *J. Chem. Phys.*, **25**, 1291.
Crawford, G. J. B., and Foster, J. S., 1956, *Canad. J. Phys.*, **34**, 653.
Curtis, A. J., McGeer, P. L., Rathmann, G. B., and Smyth, C. P., 1952, *J. Amer. Chem. Soc.*, **74**, 644.
Das, T. P., and Bersohn, R., 1956, *Phys. Rev.*, **104**, 849.
Dharmatti, S. S., and Weaver, H. E., 1952, *Phys. Rev.*, **87**, 675.
Dickinson, W. C., 1950 a, *Phys. Rev.*, **77**, 736 ; 1950 b, *Ibid.*, **80**, 563.
Diehl, P., 1957, *Helv. Phys. Acta*, **30**, 89.
Drysdale, J. J., and Phillips, W. D., 1957, *J. Amer. Chem. Soc.*, **79**, 319.
Elliott, D. F., and Shumacher, R. T., 1957, *J. Chem. Phys.*, **26**, 1350.
Feher, G., and Knight, W. D., 1955, *Rev. Sci. Instrum.*, **26**, 293.
Fermi, E., 1930, *Z. Phys.*, **60**, 320.
Frank, P. J., and Gutowsky, H. S., 1958, *Arch. des Sci.*, **11**, 215.
Freeman, R., Murray, G. R., and Richards, R. E., 1957, *Proc. Roy. Soc.* A, **242**, 455.
Giulotto, L., 1956, *Arch. des Sci.*, **9**, 8th Coll. ampere, p. 218.
Grabner, L., and Hughes, H., 1950, *Phys. Rev.*, **79**, 314, 829 ; 1951, *Ibid.*, **82**, 861.
Graham, D. M., and Waugh, J. S., 1957, *J. Chem. Phys.*, **27**, 968.
Gripenberg, J., 1956, *Acta Chem. Scand.*, **10**, 487.
Grunwald, E., Loewenstein, A., and Meiboom, S., 1956, *J. Chem. Phys.*, **25**, 382.
Gutowsky, H. S., 1958, *Ann. N.Y. Acad. Sci.*, **70**, 786.
Gutowsky, H. S., and Hoffman, C. J., 1950, *Phys. Rev.*, **80**, 110 ; 1951, *J. Chem. Phys.*, **19**, 1259 ; 1952, *Ibid.*, **20**, 200.
Gutowsky, H. S., and Holm, C. H., 1956, *J. Chem. Phys.*, **25**, 1228.
Gutowsky, H. S., Holm, C. H., Saika, A., and Williams, G. A., 1957, *J. Amer. Chem. Soc.*, **79**, 4596.
Gutowsky, H. S., and McCall, D. W., 1954, *J. Chem. Phys.*, **22**, 162.
Gutowsky, H. S., McCall, D. W., McGarvey, B. R., and Meyer, L. H., 1951, *J. Chem. Phys.*, **19**, 1328 ; 1952, *J. Amer. Chem. Soc.*, **74**, 4809.
Gutowsky, H. S., McCall, D. W., and Slichter, C. P., 1953, *J. Chem. Phys.*, **21**, 279.

GUTOWSKY, H. S., MCCALL, D. W., SLICHTER, C. P., and MCNEIL, E. R., 1951 a, *Phys. Rev.*, **82**, 748 ; 1951 b, *Ibid.*, **84**, 589.
GUTOWSKY, H. S., and MCCLURE, R. E., 1951, *Phys. Rev.*, **81**, 276.
GUTOWSKY, H. S., and MCGARVEY, B. R., 1953, *J. Chem. Phys.*, **21**, 1423.
GUTOWSKY, H. S., ROUTLEDGE, R. L., TAMRES, M., and SEARLES, S., 1954, *J. Amer. Chem. Soc.*, **76**, 4242.
GUTOWSKY, H. S., and SAIKA, A., 1953, *J. Chem. Phys.*, **21**, 1688.
GUTOWSKY, H. S., and WOESSNER, D. E., 1956, *Phys. Rev.*, **104**, 843.
HAHN, E. L., 1950, *Phys. Rev.*, **80**, 580.
HAHN, E. L., and KAPLAN, D. E., 1958, *J. Phys. Radium*, **19**, 57.
HAHN, E. L., and MAXWELL, D. E., 1951, *Phys. Rev.*, **84**, 1246 ; 1952, *Ibid.*, **88**, 1070.
HAMEKA, H. F., 1958, *Mol. Phys.*, **1**, 203.
HAZATO, G., 1957, *J. Chem. Phys.*, **27**, 605.
HEESCHEN, J. P., and GUTOWSKY, H. S., 1959, to be published.
HERZOG, B., and HAHN, E. L., 1956, *Phys. Rev.*, **103**, 148.
HOLDER, B. E., and KLEIN, M. P., 1955, *Phys. Rev.*, **98**, 265.
HOLM, C. H., 1957, *J. Chem. Phys.*, **26**, 707.
HOLZMAN, G. R., LAUTERBUR, P. C., ANDERSON, J. H., and KOTH, W., 1956, *J. Chem. Phys.*, **25**, 172.
HOOD, G. C., REDLICH, O., and REILLY, C. A., 1954, *J. Chem. Phys.*, **22**, 2067.
HORNIG, J. F., and HIRSCHFELDER, J. O., 1955, *J. Chem. Phys.*, **23**, 474.
HUGGINS, C. M., PIMENTEL, G. C., and SHOOLERY, J. N., 1955, *J. Chem. Phys.*, **23**, 1244 ; 1956, *J. Phys. Chem.*, **60**, 1311.
HUGHES, V. W., and GEIGER, J. S., 1955, *Phys. Rev.*, **99**, 1842.
HYLLERAAS, E., and SKAVLAM, S., 1950, *Phys. Rev.*, **79**, 117.
ISHIGURO, E., 1958, *Phys. Rev.*, **111**, 203.
ISHIGURO, E., and KOIDE, S., 1954, *Phys. Rev.*, **94**, 350.
ITOH, J., and YAMAGATA, Y., 1956, *J. Chem. Phys.*, **24**, 621.
JACOBSOHN, B. A., and WANGSNESS, R. K., 1948, *Phys. Rev.*, **73**, 942.
KAPLAN, J., 1958, *J. Chem. Phys.*, **28**, 278.
KAPLAN, J., and MEIBOOM, S., 1957, *Phys. Rev.*, **106**, 499.
KARPLUS, M., ANDERSON, D. H., FARRAR, T. C., and GUTOWSKY, H. S., 1957, *J. Chem. Phys.*, **27**, 597.
KIVELSON, D., and KIVELSON, M. G., 1958, *J. Mol. Spect.*, **2**, 518.
KNIGHT, W. D., 1949, *Phys. Rev.*, **76**, 1259.
KOLSKY, H. G., PHIPPS, T. E., RAMSEY, N. F., and SILSBEE, H. B., 1952, *Phys. Rev.*, **87**, 395.
KUMLER, W. D., SHOOLERY, J. N., and BOUTCHER, F. V., JR., 1958, *J. Amer. Chem. Soc.*, **80**, 2533.
KUSCH, P., 1954, *Phys. Rev.*, **93**, 1022.
LAMB, W. E., 1941, *Phys. Rev.*, **60**, 817.
LAUTERBUR, P. C., 1957, *J. Chem. Phys.*, **26**, 217 ; 1958, *Ann. N.Y. Acad. Sci.*, **70**, 841.
LEMIEUX, R. U., KULLNIG, R. K., BERNSTEIN, H. J., and SCHNEIDER, W. G., 1957, *J. Amer. Chem. Soc.*, **79**, 1005.
LEMIEUX, R. V., KULLNIG, R. K., and MOIR, R. Y., 1958, *J. Amer. Chem. Soc.*, **80**, 2237.
LIDDEL, U., and RAMSEY, N. F., 1951, *J. Chem. Phys.*, **19**, 1608.
MCCONNELL, H. M., 1956, *J. Chem. Phys.*, **24**, 460 ; 1957 a, *Ibid.*, **27**, 226 ; 1957 b, *J. Mol. Spect.*, **1**, 11.
MCCONNELL, H. M., and HOLM, C. H., 1956, *J. Chem. Phys.*, **25**, 1289.
MCCONNELL, H. M., MCLEAN, A. D., and REILLY, C. A., 1955, *J. Chem. Phys.*, **23**, 1152.
MCCONNELL, H. M., and THOMPSON, D. D., 1957, *J. Chem. Phys.*, **26**, 958.
MCGARVEY, B. R., 1957 a, *J. Chem. Phys.*, **26**, 221 ; 1957 b, *Ibid.*, **27**, 68.
MCNEIL, E. R., SLICHTER, C. P., and GUTOWSKY, H. S., 1951, *Phys. Rev.*, **84**, 1245.
MARSHALL, T. W., and POPLE, J. A., 1958, *Mol. Phys.*, **1**, 199.
MASUDA, Y., and KANDA, T., 1953, *J. Phys. Soc., Japan*, **8**, 432.
MECKE, R., 1950, *Disc. Faraday Soc.*, **9**, 161.
MEYER, L. H., and GUTOWSKY, H. S., 1953, *J. Phys. Chem.*, **57**, 481.

MEYER, L. H., SAIKA, A., and GUTOWSKY, H. S., 1953, *J. Amer. Chem. Soc.*, **75**, 4567.
MULLER, N., LAUTERBUR, P. C., and GOLDENSIN, J., 1956, *J. Amer. Chem. Soc.*, **78**, 3557.
NACHOD, F. C., 1958, *Ann. N.Y. Acad. Sci.*, **70**, 765.
NEDERBRAGT, E. W., and REILLY, C. A., 1956, *J. Chem. Phys.*, **24**, 1110.
OGG, R. A., 1954, *J. Chem. Phys.*, **22**, 560 ; 1957, *Helv. Phys. Acta*, **30**, 89.
OGG, R. A., and RAY, J. D., 1955, *Disc. Faraday Soc.*, **19**, 240 ; 1957 a, *J. Chem. Phys.*, **26**, 1339 ; 1957 b, *Ibid.*, **26**, 1515 ; 1957 c, *Ibid.*, **26**, 1340.
OVERHAUSER, A. W., 1953, *Phys. Rev.*, **91**, 476, **92**, 411.
PACKARD, M. E., and ARNOLD, J. T., 1951, *Phys. Rev.*, **83**, 210.
PACKARD, M. E., and VARIAN, R., 1954, *Phys. Rev.*, **93**, 941.
PAKE, G. E., 1948, *J. Chem. Phys.*, **16**, 327 ; 1956, *Solid State Physics*. Vol. 2 (New York : Academic Press).
PARKS, J. R., 1957, *J. Amer. Chem. Soc.*, **79**, 757.
PAULING, L., 1936, *J. Chem. Phys.*, **4**, 673.
PHILLIPS, W. D., 1958, *Ann. N.Y. Acad. Sci.*, **70**, 817.
PHILLIPS, W. D., LOONEY, C. E., and SPAETH, C. P., 1957, *J. Mol. Spect.*, **1**, 35.
PIETTE, L. H., RAY, J. D., and OGG, R. A., 1957, *J. Chem. Phys.*, **26**, 1341 ; 1958, *J. Mol. Spect.*, **2**, 66.
PIETTE, L. H., and WEAVER, H. E., 1958, *J. Chem. Phys.*, **28**, 735.
PINES, D., and SLICHTER, C. P., 1955, *Phys. Rev.*, **100**, 1014.
POPLE, J. A., 1956, *J. Chem. Phys.*, **24**, 1111 ; 1957, *Proc. Roy. Soc.* A, **239**, 541, 550 ; 1958 a, *Mol. Phys.*, **1**, 3 ; 1958 b, *Ibid.*, **1**, 168 ; 1958 c, *Ibid.*, **1**, 216.
POPLE, J. A., SCHNEIDER, W. G., and BERNSTEIN, H. J., 1957, *Canad. J. Chem.*, **35**, 1060.
PORTIS, A. M., 1956, *Phys. Rev.*, **104**, 584.
POWLES, J. G., and CUTLER, D., 1957, *Nature, Lond.*, **180**, 1344 ; 1958, *Arch. des Sci.*, **11**, 209.
PRIMAS, H., and GÜNTHARD, Hs. H., 1958, *Helv. Phys. Acta*, **31**, 413.
PROCTOR, W. G., and YU, F. C., 1951, *Phys. Rev.*, **81**, 20.
QUINN, W. E., and BROWN, R. M., 1953, *J. Chem. Phys.*, **21**, 1605.
RAMSEY, N. F., 1950 a, *Phys. Rev.*, **77**, 567 ; 1950 b, *Ibid.*, **78**, 699 ; 1951, *Ibid.*, **83**, 540 ; 1952, *Ibid.*, **86**, 243 ; 1953, *Ibid.*, **91**, 303 ; 1956, *Molecular Beams* (Oxford : University Press).
RAMSEY, N. F., and PURCELL, E. M., 1952, *Phys. Rev.*, **85**, 143.
REDFIELD, A. G., 1955, *Phys. Rev.*, **98**, 1787.
REID, C., 1956, *J. Amer. Chem. Soc.*, **78**, 3225 ; 1957, *J. Mol. Spect.*, **1**, 18.
REILLY, C. A., 1956, *J. Chem. Phys.*, **25**, 604.
REILLY, C. A., and STROMBOTNE, R. L., 1957, *J. Chem. Phys.*, **26**, 1338.
REMPEL, R. C., 1958, *Phys. Rev.*, **109**, 831.
RICHARDS, R. E., and SCHAEFER, T., 1958 a, *Trans. Faraday Soc.*, **54**, 1280 ; 1958 b, *Mol. Phys.*, **1**, 331.
ROBERTS, J. D., 1956, *J. Amer. Chem. Soc.*, **78**, 4495.
ROCARD, J. M., 1957, *Arch. des Sci.*, **10**, 209.
ROCARD, J. M., and FRANK, P. S., 1958, *C. R. Acad. Sci., Paris*, **246**, 2466.
ROUX, D., 1957, *Arch. des Sci.*, **10**, 218 ; 1958, *Helv. Phys. Acta*, **31**, 511.
SAIKA, A., and GUTOWSKY, H. S., 1956, *J. Amer. Chem. Soc.*, **78**, 4818.
SAIKA, A., and SLICHTER, C. P., 1954, *J. Chem. Phys.*, **22**, 26.
SALWEN, H., 1955, *Phys. Rev.*, **99**, 1274.
SCHIFF, L., 1955, *Quantum Mechanics*, 2nd Edn (New York : McGraw-Hill).
SCHMIDT, B. M., BROWN, L. C., and WILLIAMS, D., 1958, *J. Mol. Spect.*, **2**, 539.
SCHNEIDER, W. G., and BERNSTEIN, H. J., 1956, *Trans. Faraday Soc.*, **52**, 13.
SCHNEIDER, W. G., BERNSTEIN, H. J., and POPLE, J. A., 1958, *J. Chem. Phys.*, **28**, 601.
SCHNEIDER, W. G., and REEVES, L. W., 1958, *Ann. N.Y. Acad. Sci.*, **70**, 858.
SHOOLERY, J. N., 1954, *Anal. Chem.*, **26**, No. 9 ; 1955, *Disc. Faraday Soc.*, **19**, 215.
SHOOLERY, J. N., and CRAWFORD, B., 1957, *J. Mol. Spect.*, **1**, 270.
SOLOMON, I., 1955, *Phys. Rev.*, **99**, 559.
SOLOMON, I., and BLOEMBERGEN, N., 1956, *J. Chem. Phys.*, **25**, 261.

STEPHEN, M. J., 1957 a, *Proc. Roy. Soc.* A, **243**, 264 (Eqn (10) should read $\sigma_{\alpha\beta} = L^{-5}[L^2\chi_{\alpha\beta} - 3\chi_{\alpha\gamma}L_\gamma L_\beta]$) ; 1957 b, *Ibid.*, **243**, 274 ; 1958, *Molecular Physics*, **1**, 223.
THOMAS, H. A., 1950, *Phys. Rev.*, **80**, 901.
TIERS, G. V. D., 1956, *J. Amer. Chem. Soc.*, **78**, 2914.
TILLIEU, J., 1957, *Ann. Phys., Paris*, **2**, 471, 631.
TILLIEU, J., and GUY, J., 1956, *J. Chem. Phys.*, **24**, 1117.
TOMITA, K., 1958, *Prog. Theor. Phys.*, Japan, **19**, 541.
Varian Associates quoted with permission.
VAN VLECK, J. H., 1932, *Electric and Magnetic Susceptibilities* (Oxford : University Press), p. 276 ; 1948, *Phys. Rev.*, **74**, 1168.
VAN WAZER, J. R., CALLIS, C. F., SHOOLERY, J. N., and JONES, R. C., 1956, *J. Amer. Chem. Soc.*, **78**, 5715.
WANG, S. C., 1928, *Phys. Rev.*, **31**, 579.
WEAVER, H. E., TOLBERT, B. M., LA FORCE, R. C., 1955, *J. Chem. Phys.*, **23**, 1956.
WEINBAUM, S., 1933, *J. Chem. Phys.*, **1**, 593.
WERTZ, J. E., JAIN, P. L., and BATDORF, R. L., 1956, *Phys. Rev.*, **102**, 920.
WILLIAMS, G. A., and GUTOWSKY, H. S., 1956, *J. Chem. Phys.*, **25**, 1288.
WILLIAMS, G. A., MCCALL, D. W., and GUTOWSKY, H. S., 1954, *Phys. Rev.*, **93**, 1428.
WILLIAMS, R. B., 1958, *Ann. N.Y. Acad. Sci.*, **70**, 890.
WILSON, E. B., Jr., 1957, *J. Chem. Phys.*, **27**, 60.
WIMETT, T. F., 1953, *Phys. Rev.*, **91**, 476 (A).

EXPERIMENTAL WORK WITH ^{3}He

By V. P. PESHKOV and K. N. ZINOV'EVA
Institute for Physical Problems, Moscow

CONTENTS

		PAGE
§ 1.	Methods of separation of ^{3}He	504
§ 2.	Saturated vapour pressure. Critical point	515
§ 3.	Equation of state of gaseous ^{3}He	517
§ 4.	Thermomolecular pressure ratios for ^{3}He	521
§ 5.	Melting curve and α–β transition in solid ^{3}He	521
§ 6.	Density of liquid, gaseous and solid ^{3}He	525
§ 7.	Latent heat of vaporization	529
§ 8.	Specific heat	530
§ 9.	Entropy	532
§ 10.	Magnetic properties of ^{3}He	536
§ 11.	Absence of superfluidity	540
§ 12.	Viscosity	541
§ 13.	Thermoconductivity	545
§ 14.	Temperature discontinuity on the boundary of ^{3}He and the wall	547
§ 15.	Surface tension	547
§ 16.	Velocity of sound	548
§ 17.	Compressibility of ^{3}He	550
§ 18.	Adsorption	551
§ 19.	Cryostats with ^{3}He	552
§ 20.	Production of low temperatures by crystallization of ^{3}He	557
§ 21.	Conclusion	558
	Acknowledgments	558
	References	559

Abstract. The report describes all aspects of experimental work on ^{3}He: methods of extracting ^{3}He from a mixture of isotopes and the sources from which it is obtained, details of the most efficient methods and apparatus, experiments on saturated vapour pressure and the critical point, the equation of state of gaseous ^{3}He, thermomolecular pressure rates, melting curve and α–β transition in solid ^{3}He, experiments and data on the density of liquid, gaseous and solid ^{3}He, the thermal properties of ^{3}He—latent heat, specific heat and entropy—the magnetic properties of ^{3}He, and the phenomena of viscosity and thermal conductivity, results of experiments on the temperature discontinuities at the boundary between ^{3}He and a solid, the velocity of sound, the compressibility of ^{3}He and adsorption on activated charcoal, the construction and specifications of the work of cryostats, which enable us to obtain and maintain temperatures to 0·3°K by pumping off ^{3}He vapour.

The possibility is discussed of obtaining and maintaining temperatures below 0·3°K by adiabatic crystallization of ^{3}He. The article reviews with brief critical comments practically all published experimental work on the investigation of ^{3}He, and gives tables and graphs for all known properties of ^{3}He.

§ 1. Methods of Separation of ^{3}He

The helium isotope with mass 3, ^{3}He, occurs in its natural state as a small admixture to the basic isotope ^{4}He. Its concentration in helium extracted from the atmosphere is 10^{-4} %, and in helium obtained from wells and minerals ranges from 10^{-6} to 10^{-3} % (Aldrich and Nier 1948). No other isotopes of helium exist except radioactive ^{6}He with half-life of 0·82 sec (Holmes 1949).

Helium obtained by artificial means in atomic reactors (usually according to the reaction $^6Li + n \rightarrow {}^4He + {}^3H \rightarrow {}^4He + {}^3He + e^-$) also consists of a mixture of ^4He and ^3He isotopes, and therefore to study the properties of ^3He it is first necessary to separate it from mixture with ^4He. This problem remains also in later stages of research, since it is essential for the experimenter to be able to make mixtures and separate them again at will. MacIntyre *et al.* (1948) used the thermal diffusion method for the enrichment of ^3He. Their equipment consisted of three series-connected thermal diffusion columns of different diameters, with an overall length of about 10 m. Using electrical heaters of total power 16·6 kw, out of helium containing $1{\cdot}5 \times 10^{-5}$ % ^3He they obtained 14 cm^3 ($p = 760$ mm, $T = 0°C$) of helium, containing 0·21% ^3He, in twenty-four hours.

After it had been established that ^3He does not participate in the superfluid flow (Daunt, Probst and Johnston 1948) a series of low-temperature methods for the separation of helium isotopes was developed which are more efficient than that of thermal diffusion.

Lane, Fairbank, Aldrich and Nier (1948) proposed a means of increasing ^3He by heat flux. Such a method has been used by a number of authors: Taconis (1950) obtained enrichment of 4000 by processing 0·2 litre of liquid helium for each cycle; Soller, Fairbank and Crowell (1953) 3×10^4 by processing 0·7 litre; and Peshkov (1956) 10^5 by processing 0·75 litre. In the last case a mixture of concentration 0·2% was obtained from helium containing 10^{-6} % ^3He.

Another more reliable method is also based on the fact that ^3He does not participate in the superfluid motion. In this ^4He is extracted by a film or through a filter in the superfluid flow, due to the thermomechanical effect, i.e. utilizing the effect of thermo-osmosis. This method was suggested by Pollard and Davidson (1945) and was used by many authors (Frank 1946, Daunt *et al.* 1947 a, b, 1948, Eselson and Lazarev 1950). It gave enrichment from a few hundred times to 3×10^9 and enabled mixtures with concentrations of 10^{-3}–10^{-2} % ^3He to be obtained from naturally occurring helium. With the apparatus shown in figure 1 this method gave good results (Peshkov 1956 b).

Liquid helium was supplied continuously from a liquefier and maintained at atmospheric pressure in an outer Dewar vessel. It then entered an inner vessel through a control valve and copper capillary, being cooled to 2°K on the way. ^4He was removed from the inner vessel by thermo-osmosis through a filter ϕ into an intermediate vessel, from which helium vapour was pumped out; as a result of this a temperature $T = 2°K$ was maintained in the intermediate and inner vessels. The fraction of ^3He remaining in the inner vessel was enriched to 2×10^4 times and extracted by a central tube.

The capacity of the apparatus, which produced about 3–4 litres of liquid helium per hour, was determined by the power of the pumps which pumped out the helium from the intermediate vessel. The parameters governing the conditions of operation are given in figure 1.

Essentially, ^3He enrichment to any degree desired is not possible by thermo-osmosis. In the first place, as shown by the measurements of Dash and Taylor (1955), Eselson, Bereznyak and Kaganov (1956) and Ptoukha (1958), the λ-transition is displaced with increase in concentration into the region of low temperatures. Superfluidity in a ^3He–^4He 50% solution has already disappeared at 1·4°K.

Further, for fairly small concentrations of ^{3}He, Pomeranchuk's relation (1949) for thermo-osmosis of ^{3}He is valid

$$-SdT+\frac{dp}{\rho}-\frac{k}{m_4}d(XT)=0 \qquad \ldots\ldots(1)$$

where S is the entropy of 1 g of ^{4}He, ρ is the density, p the pressure, m_4 the mass of the ^{4}He atom, k Boltzmann's constant, and X the molar concentration of ^{3}He.

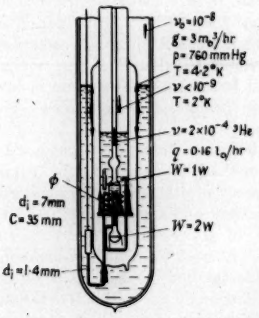


Figure 1. Apparatus for the separation of ^{3}He by thermo-osmosis and rectification.

The entropy of He II in the temperature region above 1°K may be expressed by the formula

$$S = 0{\cdot}405\ (T/2{\cdot}19)^{5{\cdot}5}\ \text{cal g}^{-1}\ \text{deg}^{-1},$$

i.e.

$$SdT = (1/6{\cdot}5)\,d(ST).$$

The density of He II changes little; therefore the complete expression can be written thus

$$d\left(\left(\frac{kX}{m_4}+\frac{S}{6{\cdot}5}\right)T-\frac{p}{\rho}\right)=0,$$

or in terms of thermo-osmosis,

$$\left(\frac{kX}{m_4}+\frac{S}{6{\cdot}5}\right)T-\frac{p}{\rho}=\text{const.} \qquad \ldots\ldots(2)$$

Since in the extraction of ^{3}He by thermo-osmosis the concentration on one side is practically equal to zero, we obtain

$$\left(\frac{kX}{m_4}+\frac{S}{6{\cdot}5}\right)T-\frac{p}{\rho}=\frac{S_0T_0}{6{\cdot}5}-\frac{p_0}{\rho}$$

or

$$X=\frac{m_4}{kT}\left(\frac{S_0T_0-ST}{6{\cdot}5}-\frac{p_0-p}{\rho}\right). \qquad \ldots\ldots(3)$$

The line of λ-transitions, and also the values of X, are calculated by the reduction formula for the differences T, plotted in the graph (figure 2). As seen in figure 2, thermo-osmosis cannot lead to any desired degree of concentration. The theoretical coefficient of separation of ^3He using superfluid filtration can be estimated from the following consideration. Assuming that superfluid flow and diffusion take place through the same effective cross section of the filter, and that the rate of the superfluid flow is limited by the critical velocity $V_s = 20$ cm sec^{-1}, then the magnitude of the superfluid flow through the unit of surface is $w_1 = \rho_s V_s$ and the flow of ^3He, as given by the diffusion, will be $w_2 = D\rho m_3 X_0/m_4 l$ where D is the coefficient of diffusion of ^3He, l is the length of the filter, ρ_s the density of the superfluid part of helium, X the molar concentration in front of the filter, m_3 and m_4 the masses of the atoms of ^3He and ^4He respectively.

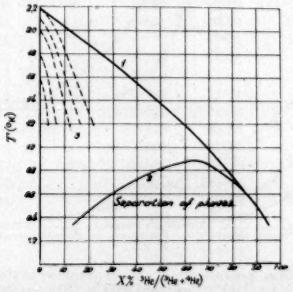


Figure 2. 1, line of λ-transitions ; 2, line of separation of phases ; 3, variation of $[(\kappa X/m_4)+(S/6{\cdot}5)]T-p/\rho$ = const. for different T_0.

The molar concentration by passing helium becomes

$$X = \frac{w_2 m_4}{w_1 m_3} = \frac{D\rho X_0}{l\rho_s V_s},$$

i.e. the maximum attainable coefficient of separation is equal to

$$A = \frac{X_0}{X} = \frac{l\rho_s V_s}{\rho D}. \qquad \ldots\ldots(4)$$

The coefficient of diffusion of ^3He in liquid helium II has been established experimentally by Beenakker *et al.* (1952). It was found to be falling rapidly from 10^{-2} cm^2 sec^{-1} at 1·3°K to 10^{-3} cm^2 sec^{-1} at 1·6°K and 4×10^{-5} cm^2 sec^{-1} at 2·1°K. With the length of the filter $l = 5$ cm, evaluation of formula (4) for the coefficient of separation for different temperatures leads to the following results : at 1·3°K, $A = 10^4$; at 1·6°K, $A = 8\times10^4$; at 2·0°K, $A = 6\times10^5$; and at 2·1°K, $A = 5\times10^5$—i.e. around 2°K the coefficient of separation is at a maximum. In practice, because of the discontinuity of the process, when diffusion continues but superfluid flow does not, and because of the imperfection of the filter, the actual

separation obtainable is less. Further, in increased concentrations part of the superfluid component decreased because of the binding of part of it by ^{3}He atoms, the volumetric velocity of the superfluid flow decreases and the efficiency of the process falls. The latter fact was verified in the work of Kuznetsov (1957). Figure 3 shows for different temperatures the dependence of the efficiency of the filter on the concentration of the fluid being enriched. Thus consideration of the thermo-osmotic method for enriching ^{3}He shows that it is always effective for the separation of ^{3}He from weak solutions, but hardly suitable for enriching solutions with concentrations higher than 5% ^{3}He.

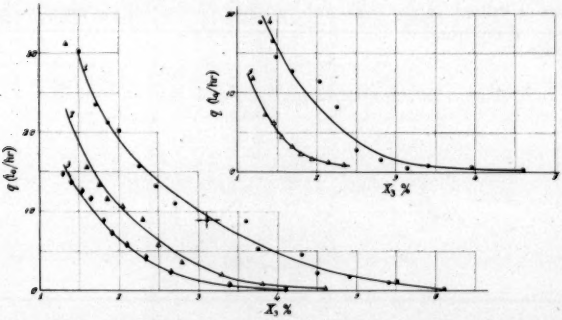


Figure 3. Dependence of filter productivity q, on concentration of the enriched liquid at different temperatures. Curve 1: 1·43°K (solution)–2·15°K (^{4}He vessel); curve 2: 1·43°–2·02°K; curve 3: 1·43°–1·86°K; curve 4: 1·86°–2·15°K; curve 5: 1·86°–2·02°K.

Solutions with high concentrations of ^{3}He can be successfully enriched by rectification (Eselson and Lazarev 1950, Kuznetsov 1957, Peshkov 1956 b, Fairbank *et al.* 1953). Rectification is shown to be particularly effective for the separation of the isotopes of helium since the difference in the vapour pressures of ^{4}He and ^{3}He is very great. The arrangement of helium isotopes (^{3}He–^{4}He) for the liquid–vapour state has been investigated by a number of authors (Eselson, Lazarev and Alexeevski 1950, Weinstock, Osborne and Abraham 1950, Sommers 1952, Eselson and Bereznyak 1956).

Extensive data have been obtained by Eselson and Bereznyak (1956). Their data, given in tables 1, 2 and 3, agree within the limits of accuracy of measurement with the results of Sommers (1952) whose measurements covered a smaller region of concentration, and with the data of Peshkov and Kachinski (1956) for higher concentrations; table 1 shows the temperature dependence of vapour pressure for different concentrations in the liquid phase, table 2 gives the dew points for vapour of different concentrations, and table 3 the distribution of ^{3}He between vapour and liquid. The curve of the liquid–vapour state for a mixture of helium isotopes under different pressures calculated by Kuznetzov (unpublished dissertation) from the data of Eselson and Bereznyak (1956) is given in figure 4.

Table 1. Temperature Dependence of Vapour Pressure for Different Compositions of the Liquid Phase X_{liq}

X_{liq} (%)	T (°K)	p (mm)	X_{liq} (%)	T (°K)	p (mm)	X_{liq} (%)	T (°K)	p (mm)
0·4	1·377	2·22	1·9	1·358	3·53	4·0	1·358	5·32
	1·387	2·35		1·377	3·75		1·358	5·31
	1·489	3·84		1·495	5·86		1·495	8·16
	1·594	6·01		1·572	7·73		1·559	9·79
	1·675	8·26		1·668	10·59		1·655	12·91
	1·755	11·12		1·743	13·33		1·745	16·56
	1·836	14·68		1·815	16·61		1·855	22·06
	1·918	19·04		1·877	19·88		1·939	27·23
	1·993	23·80		1·978	26·16		2·039	34·42
	2·070	29·36		2·037	30·26		2·120	40·89
	2·124	33·67		2·102	35·36		2·169	45·47
	2·172	38·00		2·149	39·39		2·191	48·07
	2·201	40·94		2·163	40·59		2·202	49·34
	2·301	51·56		2·177	41·91		2·254	55·06
	2·406	64·80		2·203	44·98		2·397	73·81
	2·554	87·05		2·223	47·13		2·468	84·34
	2·694	112·15		2·298	55·41		2·485	87·29
	2·793	133·11		2·424	71·83		2·549	97·93
	2·902	158·09		2·533	88·68		2·671	120·46
	3·051	197·55		2·608	101·76		2·799	148·66
	3·104	213·54		2·696	118·59		2·950	186·04
	3·202	243·60		2·740	127·95		3·105	231·36
	3·211	246·60		2·790	138·89			
				2·914	168·08			
0·8	1·358	2·46		3·073	211·98	6·3	1·358	6·96
	1·563	5·87		3·217	257·23		1·489	10·11
	1·634	7·71					1·618	14·30
	1·763	12·21					1·718	18·42
	1·815	14·50					1·828	24·08
	1·927	20·46	3·0	1·377	4·79		1·958	32·30
	2·008	25·76		1·387	4·90		2·008	35·92
	2·105	33·03		1·549	8·36		2·159	48·91
	2·126	34·85		1·572	9·23		2·183	51·49
	2·153	37·15		1·644	11·34		2·194	53·08
	2·165	38·26		1·721	14·23		2·258	60·30
	2·171	38·86		1·836	19·62		2·336	70·54
	2·175	39·17		1·977	28·23		2·428	84·13
	2·181	39·78		2·053	33·84		2·525	100·11
	2·188	40·74		2·093	36·94		2·603	114·57
	2·208	42·83		2·102	37·80		2·665	126·88
	2·248	46·89		2·144	41·16		2·701	134·60
	2·380	62·66		2·175	44·10			
	2·449	72·32		2·197	46·72			
	2·548	87·58		2·249	52·30	8·3	1·352	8·10
	2·748	125·10		2·348	64·50		1·358	8·88
	2·815	139·69		2·460	80·45		1·483	12·55
	2·900	159·38		2·513	88·93		1·598	16·77
	2·902	160·04		2·635	110·63		1·696	21·05
	3·003	186·35		2·752	134·73		1·851	29·48
	3·050	199·43		2·850	157·33		1·983	38·52
	3·052	199·82		2·979	190·23		2·085	46·53
	3·167	234·73		3·124	233·17		2·107	48·71
	3·202	245·97		3·197	256·64			

Table 1. *Continued.*

X_{liq} (%)	T (°K)	p (mm)	X_{liq} (%)	T (°K)	p (mm)	X_{liq} (%)	T (°K)	p (mm)
8·3	2·144	52·49	13·4	2·409	103·12	19·2	2·469	129·86
	2·172	55·67		2·501	119·52		2·571	152·18
	2·187	58·40		2·600	140·17		2·688	180·74
	2·207	60·45		2·701	163·41		2·799	211·32
	2·248	65·67		2·802	189·58		2·900	241·21
	2·402	87·77		2·884	212·07		2·999	273·69
	2·565	116·39		3·063	267·83		3·107	312·17
	2·720	148·91		3·215	321·79		3·202	348·09
	2·881	189·59						
	3·023	230·39	16·7	1·388	14·86	22·6	1·336	15·14
	3·206	291·15		1·463	17·84		1·353	15·68
	3·219	296·01		1·496	19·23		1·460	20·74
				1·596	23·86		1·502	22·80
11·1	1·354	10·32		1·701	29·80		1·554	25·64
	1·363	10·94		1·771	34·01		1·646	30·52
	1·461	14·19		1·849	39·14		1·704	33·99
	1·467	14·58		1·896	42·53		1·802	40·58
	1·542	17·27		1·930	45·06		1·831	42·53
	1·635	21·32		1·955	47·15		1·864	45·14
	1·730	26·16		1·980	49·79		1·905	48·88
	1·842	32·96		2·037	55·60		1·968	55·53
	1·936	39·42		2·107	63·35		2·007	59·75
	2·015	44·96		2·138	67·25		2·068	66·98
	2·020	45·62		2·174	71·79		2·101	71·32
	2·086	51·88		2·189	75·00		2·126	74·71
	2·125	56·13		2·213	78·15		2·173	81·47
	2·148	58·71		2·298	90·91		2·202	86·68
	2·180	62·49		2·384	106·13		2·294	101·82
	2·200	65·97		2·530	134·41		2·414	124·29
	2·298	79·37		2·656	162·07		2·474	136·54
	2·476	108·64		2·805	200·80		2·552	154·02
	2·699	154·57		2·900	228·16		2·611	168·17
	2·796	178·77		3·028	269·00		2·719	196·36
	2·904	207·62		3·155	314·27		2·809	222·64
	3·023	243·69		3·219	339·33		2·906	252·03
	3·215	309·33					3·052	301·99
			19·2	1·360	14·61		3·136	333·87
13·4	1·342	11·63		1·379	15·59		3·211	363·31
	1·399	12·74		1·474	19·19			
	1·469	15·99		1·558	23·60	23·9	1·336	15·54
	1·553	19·64		1·650	28·57		1·362	16·39
	1·628	23·15		1·706	32·10		1·406	18·21
	1·725	28·48		1·752	35·06		1·498	22·65
	1·807	33·34		1·839	41·04		1·579	27·15
	1·844	35·89		1·883	44·91		1·710	35·07
	1·944	42·87		1·930	48·58		1·806	41·82
	1·995	46·88		1·974	53·62		1·826	43·10
	2·045	51·43		2·027	59·13		1·873	47·81
	2·087	56·30		2·081	65·43		1·908	51·12
	2·122	60·15		2·128	71·07		1·948	54·97
	2·156	64·07		2·175	77·42		2·038	65·46
	2·181	67·19		2·233	87·67		2·149	80·07
	2·199	70·59		2·278	94·66		2·184	85·28
	2·301	85·25		2·376	111·85		2·198	88·61

Table 1. *Continued.*

X_{liq} (%)	T (°K)	p (mm)	X_{liq} (%)	T (°K)	p (mm)	X_{liq} (%)	T (°K)	p (mm)
23·9	2·279	100·87	38·3	2·170	107·71	73·4	1·515	38·52
	2·458	138·48		2·194	113·66		1·596	48·30
	2·601	171·54		2·199	114·78		1·691	61·06
	2·754	213·45		2·253	125·72		1·806	79·06
	2·885	253·67		2·306	137·19		1·897	95·09
	3·035	305·70		2·331	142·88		2·007	117·08
	3·206	372·35		2·408	161·34		2·108	140·19
				2·488	181·52		2·152	150·56
30·2	1·316	16·23		2·554	199·62		2·177	157·95
	1·341	17·11		2·604	213·13		2·194	162·98
	1·380	18·70		2·680	237·15		2·257	181·08
	1·423	21·59		2·757	262·36		2·306	195·48
	1·485	24·60		2·856	296·94		2·404	226·82
	1·537	27·80					2·499	259·30
	1·597	31·24	52·7	1·350	21·06		2·606	298·89
	1·649	34·40		1·356	21·83		2·710	340·66
	1·722	39·73		1·437	26·70			
	1·735	40·71		1·471	28·94	82·4	1·380	28·56
	1·784	45·58		1·519	34·15		1·392	30·15
	1·830	49·96		1·562	36·58		1·490	41·09
	1·911	58·94		1·592	40·29		1·595	53·71
	1·972	66·74		1·652	45·84		1·694	68·86
	1·992	69·14		1·686	49·87		1·808	88·66
	2·069	79·85		1·728	54·79		1·894	105·38
	2·126	88·03		1·803	64·71		1·996	128·22
	2·170	95·37		1·877	75·24		2·085	149·67
	2·202	102·22		1·944	85·94		2·176	172·34
	2·218	105·97		2·009	97·14		2·199	182·17
	2·302	121·12		2·054	105·32		2·256	199·37
	2·375	135·95		2·106	115·41		2·308	216·63
	2·508	166·50		2·163	127·30		2·400	248·31
	2·655	205·91		2·193	136·18		2·504	287·36
	2·773	241·97		2·254	150·86		2·603	328·42
	2·878	276·70		2·330	171·00		2·695	369·74
	3·004	322·59		2·390	186·91			
	3·102	361·65		2·474	211·68			
	3·205	404·75		2·540	232·01	90·8	1·392	31·26
				2·617	257·47		1·405	32·49
38·3	1·331	18·99		2·669	275·95		1·507	45·32
	1·334	18·95		2·737	301·49		1·601	58·41
	1·389	21·76		2·786	320·42		1·678	70·71
	1·422	23·21					1·809	94·55
	1·471	25·80					1·906	115·67
	1·543	30·11	56·3	2·255	157·26		2·004	138·61
	1·570	32·27		2·352	181·33		2·097	163·04
	1·641	37·91		2·448	208·82		2·139	175·30
	1·715	44·94		2·535	235·80		2·182	186·34
	1·757	49·06		2·629	267·32		2·193	192·94
	1·804	53·89		2·703	294·96		2·259	214·74
	1·857	60·38		2·774	322·21		2·303	229·99
	1·949	72·69					2·395	263·90
	2·046	86·79	73·4	1·400	27·27		2·507	309·00
	2·090	94·65		1·411	26·52		2·586	343·55
	2·141	102·24		1·422	26·89		2·648	371·96

Table 2. Temperature Dependence of Pressure at the Beginning of Condensation for Mixtures of Different ^{3}He Content

X_{vap} (%)	T (°K)	p (mm)	X_{vap} (%)	T (°K)	p (mm)	X_{vap} (%)	T (°K)	p (mm)
1·9	2·037	26·4	29·0	1·986	31·5	57·6	1·998	51·4
	2·166	37·2		2·349	78·4		2·101	66·5
	2·369	59·7		2·533	112·7		2·444	137·4
	2·520	81·5		2·676	146·3		2·652	204·4
	2·776	129·3		2·826	187·0		2·825	267·4
	3·033	193·5		3·012	248·3			
						73·4	1·451	9·5
4·0	1·748	10·4	35·4	1·404	3·3		1·691	29·6
	2·077	30·2		1·836	21·9		1·865	51·3
	2·411	66·4		2·044	40·2		2·254	127·1
	2·732	122·3		2·159	54·4		2·489	199·3
	3·040	198·7		2·298	76·4		2·731	294·8
				2·516	121·1			
11·6	1·350	1·9		2·706	167·6	82·4	1·412	16·1
	1·795	14·0		2·808	197·0		1·611	30·3
	2·131	38·0		2·899	225·7		1·818	59·0
	2·309	58·7		3·100	299·7		2·106	119·7
	2·628	111·8		3·355	418·6		2·283	173·7
	2·853	164·2					2·404	210·5
	3·229	281·6	57·6	1·433	5·7		2·640	296·3
				1·581	13·8			
29·0	1·410	3·3		1·811	31·5			
	1·768	15·5		1·963	47·8			

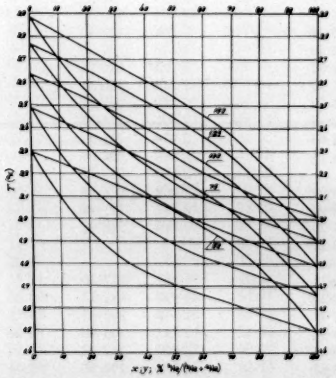


Figure 4. Liquid–vapour phase diagram for the system of helium isotopes at different vapour pressures. The numbers on the curves show pressure in mm Hg.

Table 3. Distribution of ^3He between Vapour and Liquid (X_{liq}/X_{vap}) for Different Compositions of the Liquid Phase X_{liq} and for Different Temperatures

	X_{liq}/X_{vap}									
X_{liq} (%)	1·4°K	1·5°K	1·6°K	1·7°K	1·8°K	1·9°K	2·0°K	2·1°K	2·2°K	2·3°K
0·004	39·75	29·25	22·25	16·75	13·00	10·25	8·75	7·50	6·75	6.00
0·008	33·75	25·62	20·00	15·87	12·62	10·12	8·37	6·87	6·00	5·37
0·019	24·53	19·63	16·05	13·26	11·00	9·21	7·79	6·68	5·74	5·05
0·030	19·17	16·03	13·53	11·57	9·93	8·43	7·10	6·00	5·10	4·43
0·040	16·10	13·82	11·92	10·22	8·67	7·47	6·42	5·50	4·80	4·25
0·063	11·67	10·43	9·24	8·11	7·11	6·35	5·75	5·21	4·59	3·98
0·083	9·34	8·52	7·76	7·04	6·24	5·70	5·22	4·77	4·30	3·87
0·111	7·32	6·76	6·29	5·79	5·33	4·90	4·57	4·19	3·85	3·61
0·134	6·22	5·79	5·42	5·04	4·68	4·36	4·07	3·79	3·51	3·33
0·167	5·12	4·82	4·53	4·24	3·96	3·74	3·53	3·32	3·12	2·99
0·192	4·53	4·28	4·03	3·80	3·58	3·39	3·22	3·04	2·87	2·79
0·226	3·91	3·71	3·52	3·34	3·15	3·01	2·88	2·73	2·61	2·54
0·239	3·72	3·53	3·36	3·19	3·02	2·89	2·76	2·64	2·53	2·46
0·302	3·00	2·87	2·75	2·64	2·53	2·44	2·35	2·27	2·20	2·15
0·383	2·39	2·31	2·24	2·17	2·11	2·05	1·99	1·94	1·90	1·86
0·527	1·78	1·73	1·69	1·66	1·63	1·61	1·58	1·55	1·53	1·51
0·563	—	—	—	—	—	—	—	—	1·46	1·44
0·734	1·31	1·29	1·28	1·26	1·26	1·24	1·23	1·22	1·21	1·20
0·824	1·18	1·17	1·16	1·15	1·15	1·14	1·14	1·13	1·12	1·12
0·908	1·09	1·08	1·08	1·07	1·07	1·07	1·06	1·06	1·06	1·06

	X_{liq}/X_{vap}								
X_{liq} (%)	2·4°K	2·5°K	2·6°K	2·7°K	2·8°K	2·9°K	3·0°K	3·1°K	3·2°K
0·004	5·25	4·75	4·25	3·75	3·50	3·00	2·75	2·62	2·50
0·008	4·75	4·37	4·00	3·62	3·25	3·00	2·75	2·62	2·50
0·019	4·42	4·05	3·68	3·37	3·05	2·84	2·63	2·47	2·26
0·030	4·03	3·70	3·30	3·10	2·87	2·67	2·47	2·27	2·07
0·040	3·97	3·62	3·20	3·07	2·82	2·57	2·40	2·22	—
0·063	3·75	3·44	3·17	3·00	—	—	—	—	—
0·083	3·54	3·25	3·04	2·89	2·73	2·54	2·37	2·20	2·06
0·111	3·31	3·05	2·86	2·71	2·60	2·45	2·31	2·17	2·04
0·134	3·11	2·90	2·72	2·59	2·48	2·36	2·25	2·14	2·04
0·167	2·84	2·66	2·52	2·43	2·33	2·19	2·10	—	—
0·192	2·66	2·50	2·39	2·33	2·23	—	—	—	—
0·226	2·44	2·33	2·24	2·19	2·10	—	—	—	—
0·239	2·37	2·27	2·19	2·14	2·05	—	—	—	—
0·302	2·08	2·03	1·96	1·93	1·87	—	—	—	—
0·383	1·81	1·78	1·73	1·71	1·68	—	—	—	—
0·527	1·48	1·47	1·44	1·43	1·41	—	—	—	—
0·563	1·42	1·41	1·39	1·37	1·36	—	—	—	—
0·734	1·20	1·19	1·18	1·17	—	—	—	—	—
0·824	1·12	1·11	1·11	1·10	—	—	—	—	—
0·908	1·05	1·05	1·05	1·05	—	—	—	—	—

As mentioned above, rectification seems to be a very efficient method for the separation of the isotopes of helium. An estimate of the action of the rectification column, as given in the work of Peshkov (1956 a), leads to the following results.

Distribution of concentration of ^{4}He in the liquid phase along the column in the non-extraction region is given by the formula

$$\frac{X}{(1-X)^{1/\alpha}} = \frac{X_0}{(1-X_0)^{1/\alpha}} e^{\beta z} \qquad \ldots\ldots(5)$$

under the condition $X_a/(1-X_a) = \alpha Y_a/(1-Y_a)$ or, in the case of extraction and $X_a = \alpha Y_a$, which for helium in the temperature interval 1·5° to 1·9°K is valid up to $X = 0{\cdot}5$ at $\alpha = 3$,

$$X = X_0 \frac{(\alpha-1)e^{\beta z} - \alpha\gamma}{\alpha - 1 - \alpha\gamma}. \qquad \ldots\ldots(6)$$

In these expressions at the region of turbulence in the column

$$\beta = \frac{\alpha(1-\gamma)-1}{12\alpha(1-\gamma)a}\left(\frac{D_v}{av}\right)^{0{\cdot}2}. \qquad \ldots\ldots(7)$$

The hold-up in the column is given by formula

$$q = pX_1/\beta l. \qquad \ldots\ldots(8)$$

If gas with high concentration of ^{3}He is extracted, the effectiveness of the column is greater, $e^{\beta l} \gg 1$; the mean concentration of the selected gas would be

$$X = \frac{[1-\alpha\gamma/(\alpha-1)]\,Q\,e^{-\beta l}}{\alpha p_0} \ln\frac{X_2}{X_1}. \qquad \ldots\ldots(9)$$

The time needed to reach a stable condition in the column, after which extraction can begin, will be

$$\tau = \frac{4\alpha p}{(\alpha-1)W}. \qquad \ldots\ldots(10)$$

In the above formulae the following symbols have been used: X is the concentration of ^{4}He in the liquid phase in the column; X_0, X_1, X_a, X_1, X_2 are the concentrations of ^{4}He in the liquid phase at the top of the column, at the base of the column, at the boundary between the liquid and vapour phases, and in the vaporizer at the beginning and end of separation, respectively; Y_a is the concentration of ^{4}He in the vapour phase at the boundary of the liquid phase; z is the distance from the top of the column; γ the fraction taken from the vapour jet at the top of the column; D_v the diffusion coefficient in the vapour phase; v_0 the mean linear velocity of the vapour jet; a the radius of the tubular column or the effective radius, approximately equal to the radius of the rings lining the column; q the number of moles of ^{4}He in the column; p the average number of moles of helium in the column; l the length of the column; Q the number of moles of ^{4}He in the column and the vaporizer; p_0 the number of helium moles removed from the column; W the velocity of the vapour jet in the column (mol sec^{-1}).

An analysis of the working of the column also leads to the conclusion that it is more convenient to work with maximum speeds of the jet just before flooding.

The combined apparatus shown in figure 5 was found to be effective for separation of ^3He from mixtures of concentrations higher than 0·1%. It consists of the rectification column 1 and filter 2. The column was made from a stainless steel tube with internal diameter 9·6 mm, external diameter 10 mm and length 200 mm, and was lined with a small ring 1·5 mm diameter made of constantan wire of diameter 0·2 mm. The filter itself was made from a copper pipe, internal diameter 2 mm and external diameter 4 mm and 45 mm long, tightly plugged by emery powder. As a rule the apparatus worked in two cycles, the extraction cycle and the enrichment cycle. The performance and data characteristics of the extraction cycle are given in figure 5 on the right of the apparatus. Helium was condensed in a coil at the rate of 100 l_0/hour (i.e. litres at 760 mm Hg and 0°c),

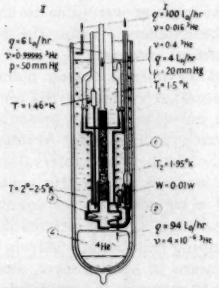


Figure 5. Apparatus for the concentration of ^3He.

and transferred to a container 3 from which ^4He was drawn off by means of thermo-osmosis, into a vessel 4. Simultaneously, a mixture enriched by rectification to a concentration of the order of 0·4 ^3He, was extracted from the upper part of the column. In the second cycle, the data for which are given in figure 5 on the left of the apparatus, a mixture of initial concentration of about 0·4 ^3He, was condensed in the vessel 3 and was then extracted from the upper part of the column, at the rate of about 10 l_0/hour. During extraction the temperature of the vessel 3 was raised from 2° to 2·5°K and the rate of extraction decreased to 4 l_0/hour. In one of the experiments, the mean concentration of the 4 l_0 of extracted gas was 0·99995 ^3He. The unprocessed 0·8 l_0 of mixture left in the column had a ^3He content of 0·25 l_0. For this apparatus $e^{\beta l} = 4000$, i.e. $1/\beta = 2{\cdot}5$ cm and the time constant $\tau_0 = 10$ sec. The apparatus described above enables ^3He to be obtained to any required degree of purity, and allows solutions of ^3He with ^4He to be used.

§2. Saturated Vapour Pressure. Critical Point

Investigation of the properties of pure ^3He began about ten years ago, when total supplies of ^3He were estimated in tenths of cubic centimetres of gas at S.T.P. Interest in this liquid was so great that in spite of the negligible quantities of ^3He available, attempts were made to investigate its basic properties.

^{3}He was first liquefied in 1948 by Sydoriak, Grilly and Hammel (1949) who had at their disposal a total of 20 cm^3 of ^{3}He at S.T.P. ^{3}He was condensed in a steel capillary with an internal diameter of 1·2 mm, immersed in a bath of liquid helium to a depth of 5–10 mm. The other end of the capillary outside the cryostat was connected to a mercury manometer and a Toepler pump. By moving the mercury piston in the pump the pressure of ^{3}He was increased in the capillary and the manometer. The cessation of pressure increase with a further decrease in volume of the system was taken as an indication of the beginning of condensation. Despite the predictions of many authors (London and Rice 1948, Tisza 1948) who doubted the possibility of liquefying ^{3}He because of its high zero point energy, Sydoriak, Grilly and Hammel obtained typical van der Waals curves for a liquid on the ^{3}He phase diagram. Furthermore, having replaced the steel capillary by a glass one, the authors were able to check visually that liquid was found in the bottom of the capillary. From the form of the isotherm it was possible to determine the critical constants $T_c = 3{\cdot}34°\text{K}$, $p_c = 875$ mm Hg. From this the authors determined the normal boiling point of ^{3}He as 3·20°K, and the curve of pressure of the saturated vapour in the temperature interval from 1·2° to the critical point. The curve of vapour pressure and the critical constants were, however, more accurately measured by Abraham, Osborne and Weinstock (1950), who investigated the temperature interval from 1·0°K to 3·35°K. The authors had 30 cm^3 of ^{3}He at S.T.P. with a ^{4}He content $0{\cdot}03 \pm 0{\cdot}03\%$. ^{3}He was condensed in the cavity of a copper block of volume 50 mm^3 connected to an outer vessel by a Monel metal tube, with an internal diameter 0·5 mm. The block with ^{3}He was cooled by pumping helium vapour from the cryostat. By regulating the speed of pumping by means of a diaphragm, situated at the entrance of the pump, the authors were able to maintain the temperature in the region of investigation accurate to 0·001°K. The vapour pressure curve according to the data obtained from this work, is well described by the empirical formula

$$\log_{10} p_{\text{mm}} = \frac{0{\cdot}97796}{T} + 2{\cdot}5 \log_{10} T + 0{\cdot}00302 T^3 + 1{\cdot}91594 \qquad \ldots\ldots(11)$$

$$1{\cdot}0° < T < 3{\cdot}3°\text{K}$$

where T is defined according to the temperature scale agreed to in 1948 (Van Dijk and Shoenberg 1949, with corrections to Kistemaker (1946)). From this equation the normal boiling-point temperature is calculated as $T = 3{\cdot}195 \pm 0{\cdot}001°\text{K}$. The critical constants derived from the data obtained in this work are $T_c = 3{\cdot}35 \pm 0{\cdot}02°\text{K}$ and $p_c = 890 \pm 20$ mm Hg. Saturated vapour pressure in the temperature region below 1°K (from 0·45 to 1·0°K) have been measured with great accuracy by Sydoriak and Roberts (1957), who pump ^{3}He vapour to cool the liquid. The apparatus of Sydoriak and Roberts is shown in figure 6. ^{3}He was condensed in a copper sphere with internal diameter 1·2 cm suspended on a thin-walled tube (70 Cu, 30 Ni) of diameter 3 mm in a copper vacuum jacket. The upper end of the tube was fastened to a brass block, as shown in the diagram, extending to the outside of the ^{3}He pump. The temperature of the liquid was measured by the magnetic susceptibility of ferric ammonium or chromopotassium alum which was placed inside the sphere. The salt was previously calibrated against the vapour pressure of ^{3}He for temperatures above 1°K. The vapour pressure was

measured with great accuracy by mercury and oil manometers. The curve of saturated vapour pressure below 1°K, according to the measurements of Sydoriak and Roberts, can be represented by the equation

$$\log_{10} p_{mm} = \frac{1{\cdot}10371}{T} + 2{\cdot}3214 \log_{10} T + 2{\cdot}0936 - 0{\cdot}08976T + 0{\cdot}03756T^2 - 0{\cdot}004T^3$$
$$0{\cdot}45° < T < 1{\cdot}0°K \qquad \ldots\ldots(12)$$

where T is defined according to the E_{55} scale (Clement, Logan and Gaffney 1955).

As well as the saturated vapour pressure, Sydoriak and Roberts also determined the temperature and pressure of the critical point, finding $T_c = 3{\cdot}329°K$,

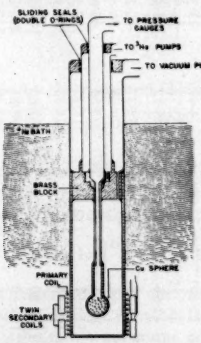


Figure 6. Sydoriak and Roberts' (1957) system for a cryostat with ^3He.

$p_c = 875$ mm Hg. It may be noted that Peshkov (1957), from the measurements of density of ^3He by an optical method, obtained for the critical point the values $T_c = 3{\cdot}38 \pm 0{\cdot}03°$, $p_c = 930 \pm 20$ mm Hg, which differs little from the preceding data.

The empirical formula (12) of Sydoriak and Roberts for saturated vapour pressure below 1°K is in good agreement with the equation of Abraham, Osborne and Weinstock (11) for higher temperatures. Table 4 gives the saturated vapour pressure values for ^3He in the interval 0·3–3·34°K, derived from the data of Sydoriak and Roberts (1957), and Abraham, Osborne and Weinstock (1950). Temperature is defined according to the scale E_{55}. It can readily be seen that the vapour pressure of ^3He in the region of low temperatures is many times greater than the corresponding pressure of ^4He. This makes it possible to use ^3He successfully as a working substance in condensation thermometers for measurement of temperatures below 1°K.

§ 3. The Equation of the Gaseous State of ^3He

(P, V) isotherms of ^3He gas (with 99·75% purity) were obtained experimentally by Keller (1955), for five temperatures in the interval 1·5–3·8°K. The second

Table 4. ^{3}He Vapour Pressures in mm Hg at 0°C at intervals of 0·01°K

T (°K)	0·00	0·01	0·02	0·03	0·04	0·05	0·06	0·07	0·08	0·09
0·3	0·00150	0·00213	0·00296	0·00404	0·00542	0·00717	0·00935	0·01205	0·01533	0·01930
0·4	0·02405	0·02969	0·03634	0·04411	0·05314	0·06355	0·07551	0·08915	0·10462	0·12210
0·5	0·1418	0·1638	0·1883	0·2155	0·2456	0·2789	0·3154	0·3554	0·3991	0·4468
0·6	0·4985	0·5546	0·6152	0·6806	0·7510	0·8267	0·9077	0·9945	1·0872	1·1860
0·7	1·291	1·403	1·522	1·647	1·780	1·921	2·069	2·225	2·390	2·563
0·8	2·744	2·935	3·135	3·344	3·563	3·792	4·031	4·280	4·540	4·810
0·9	5·092	5·385	5·689	6·005	6·333	6·673	7·026	7·391	7·769	8·160
1·0	8·564	8·981	9·413	9·858	10·318	10·792	11·280	11·783	12·301	12·835
1·1	13·384	13·948	14·528	15·124	15·737	16·366	17·012	17·674	18·354	19·051
1·2	19·765	20·497	21·247	22·015	22·801	23·606	24·429	25·271	26·132	27·013
1·3	27·913	28·833	29·772	30·732	31·711	32·711	33·732	34·774	35·837	36·920
1·4	38·03	39·15	40·30	41·47	42·66	43·88	45·12	46·38	47·66	48·97
1·5	50·30	51·65	53·03	54·43	55·85	57·30	58·77	60·27	61·79	63·34
1·6	64·91	66·51	68·13	69·78	71·46	73·16	74·88	76·64	78·42	80·23
1·7	82·06	83·92	85·81	87·73	89·67	91·64	93·64	95·67	97·73	99·81
1·8	101·93	104·07	106·24	108·44	110·67	112·94	115·23	117·55	119·90	122·28
1·9	124·69	127·14	129·61	132·12	134·66	137·23	139·83	142·46	145·12	147·82
2·0	150·55	153·31	156·11	158·94	161·80	164·70	167·63	170·59	173·59	176·62
2·1	179·68	182·78	185·92	189·09	192·30	195·54	198·81	202·13	205·47	208·86
2·2	212·28	215·74	219·23	222·76	226·33	229·93	233·58	237·26	240·97	244·73
2·3	248·52	252·36	256·23	260·14	264·08	268·07	272·10	276·17	280·27	284·42
2·4	288·60	292·83	297·10	301·41	305·75	310·14	314·57	319·04	323·56	328·11
2·5	332·71	337·35	342·03	346·75	351·52	356·33	361·18	366·08	371·02	376·00
2·6	381·02	386·09	391·21	396·37	401·57	406·82	412·11	417·45	422·83	428·26
2·7	433·73	439·25	444·81	450·42	456·08	461·78	467·53	473·33	479·17	485·06
2·8	491·00	496·99	503·02	509·10	515·23	521·40	527·63	533·90	540·22	546·59
2·9	553·01	559·48	566·00	572·56	579·18	585·84	592·56	599·32	606·13	613·00
3·0	619·92	626·89	633·90	640·97	648·09	655·26	662·48	669·76	677·08	684·46
3·1	691·88	699·36	706·89	714·48	722·11	729·80	737·54	745·34	753·18	761·08
3·2	769·04	777·04	785·10	793·21	801·38	809·60	817·87	826·20	834·58	843·01
3·3	851·50	860·05	868·64	877·30	886·00					

virial coefficient B calculated from these measurements is defined by the equation

$$\frac{PV}{N} = a + b\left(\frac{N}{V}\right) + c\left(\frac{N}{V}\right)^2 + \ldots$$

$$a = A = \boldsymbol{R}T, \quad b = aB, \quad c = aC \qquad \ldots\ldots(13)$$

shown in table 5, where the temperatures are given in the Kistemaker (1946) scale.

Table 5. Temperature Dependence of the Second Virial Coefficient B

T (°K)	B
3·782	− 65·4
2·993	− 86·3
2·150	−117·9
1·817	−142·3
1·516	−168

These data are well described by the empirical formula

$$B(\text{cm}^3\,\text{mol}^{-1}) = 10{\cdot}26 - \frac{299{\cdot}90}{T} + \frac{44{\cdot}27}{T^2}. \qquad \ldots\ldots(14)$$

The second virial coefficient, as found experimentally by Keller, agrees sufficiently closely with the values calculated in the work (de Boer *et al.* 1950, Kilpatrick *et al.* 1954, 1955) based on the Lennard-Jones potential

$$V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right].$$

The equation of state of gaseous ^3He was also derived by Peshkov (1957), who measured the density and the coefficient of isothermal compressibility of gaseous ^3He by an optical method. Peshkov's apparatus consisted of a rectangular glass chamber, cemented from plane-parallel plates, and filled with ^3He. When the chamber was illuminated by monochromatic light, bands of constant optical width could be seen on its surface. For the determination of the coefficient of isothermal compressibility, the temperature of the bath was maintained constant, and shifting of the bands with change of pressure inside the chamber was observed.

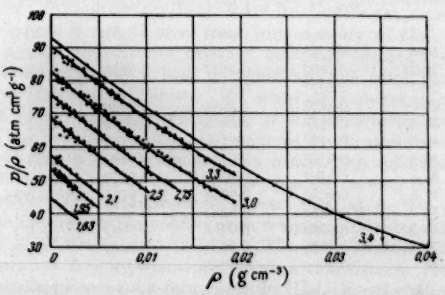


Figure 7. Compressibility curves of gaseous ^3He. The temperature is shown in °K on the curves.

The results obtained are given in figure 7, where for comparison Keller's data for 3°K are shown by crosses (at $\rho = 0{\cdot}005$, $p/\rho = 72$; $\rho = 0{\cdot}01$, $p/\rho = 60$). Keller's results for other temperatures, not shown in the diagram, are also in good agreement with Peshkov's measurements. The continuous lines in figure 7 represent the curves calculated by Peshkov from the formula

$$p/\rho = 27{\cdot}35T - 2{\cdot}3 \times 10^3\rho + 1{\cdot}8 \times 10^4\rho^2. \qquad \ldots\ldots(15)$$

As can be seen from the diagram, the experimental results are fitted well by these curves, representing the equation of state of gaseous ^{3}He. As distinct from Keller's equation, (14), in Peshkov's equation, (15), the second virial coefficient B is determined within the limits of experimental error by the relation

$$B = -243/T. \qquad \ldots\ldots(16)$$

Table 6. Experimental Values for ^{3}He Thermomolecular Pressure Ratios

(1)	(2)	(3)	(4)	(5)
	Run I : $T_c = 2{\cdot}15°K$; $T_w = 301{\cdot}2°K$			
4·87	0·4730	0·4783	0·4719	−0·2
8·12	0·5917	0·5939	0·5857	−1·0
10·94	0·6608	0·6683	0·6619	+0·2
13·67	0·7096	0·7155	0·7101	+0·1
18·97	0·7781	0·7834	0·7794	+0·2
24·2	0·8225	0·8206	0·8174	−0·6
29·0	0·8527	0·8573	0·8546	+0·2
33·9	0·8756	0·8804	0·8781	+0·3
43·8	0·9074	0·9079	0·9061	−0·1
53·1	0·9272	0·9347	0·9332	+0·6
63·0	0·9419	0·9448	0·9435	+0·2
89·2	0·9648	0·9654	0·9645	0·0
115·3	0·9761	0·9758	0·9752	−0·1
167·2	0·9870	0·9832	0·9828	−0·4
291	0·9950	0·9944	0·9943	−0·1
409	0·9973	(0·9973)	(0·9973)	
	Run II : $T_c = 1{\cdot}90°K$; $T_w = 305{\cdot}7°K$			
8·56	0·6000	0·6004	0·5955	−0·8
14·8	0·7235	0·7142	0·7095	−2·0
20·25	0·7897	0·7967	0·7930	+0·4
24·7	0·8227	0·8370	0·8338	+1·3
29·0	0·8492	0·8646	0·8620	+1·5
36·4	0·8820	0·8953	0·8932	+1·3
46·8	0·9122	0·9196	0·9179	+0·6
55·5	0·9303	0·9414	0·9400	+1·0
66·4	0·9457	0·9482	0·9470	+0·1
95·6	0·9682	0·9734	0·9726	+0·5
155	0·9846	0·9842	0·9836	−0·1
244	0·9930	0·9908	0·9903	−0·3
390	0·9971	0·9934	0·9933	−0·4
646	0·9990	(0·9990)	(0·9990)	

(1) Rp_w (cm μ Hg) ; (2) $(p_c/p_w)_{calc}$; (3) $(p_c/p_w)_{obs}$ uncorrected for adsorption ; (4) $(p_c/p_w)_{obs}$ corrected for adsorption ; (5) [(4)–(2)]/(2) (%).

§4. Thermomolecular Pressure Ratios for ^3He

Thermomolecular pressure ratios for ^3He were measured over a wide temperature interval, from room to helium temperature, by Sydoriak and Roberts (1956). The results of the measurements for the pressure ratio p_{cold} to p_{warm} are given in table 6, where the product Rp_w is expressed in cm μ Hg (R is the radius of the tube in centimetres and p_w the pressure at the warm end of the tube in microns of mercury).

§5. The Melting Curve and α–β Transitions in Solid ^3He

Experiments with ^3He at very low temperatures show that ^3He, like ^4He, cannot be transformed into the solid state by cooling at saturated vapour pressure. To obtain solid ^3He it is necessary to provide pressure of a few tens of atmospheres.

Solid ^3He was first obtained by Osborne, Abraham and Weinstock in 1951. These authors investigated the melting curve in the temperature interval 1·02–1·51°K using the block capillary method. The diagram of the equipment is shown in figure 8. The authors had 190 cm^3 of gaseous ^3He at their disposal.

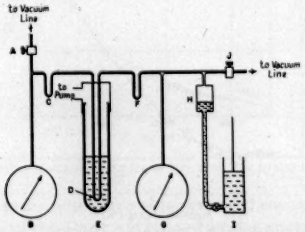


Figure 8. Apparatus for the determination of the solidifying curve (diagrammatic representation).

To decrease the volume of the system, the U-tube D immersed in a bath of liquid helium and the connecting tubes were made from steel capillaries of very small internal diameter: 0·1 mm for the U-tube and 0·5 mm for the connecting tubes. Bourdon manometers B and G were filled with mercury to minimize the free volume. To prevent the U-tube capillary becoming blocked with mercury, traps C and F immersed in a bath of liquid nitrogen were provided. The measurement process was as follows. By means of a mercury piston H the pressure in the system was increased steadily at a given constant temperature of the cryostat. The pressure increase was registered by the manometers B and G. While the liquid ^3He in the U-tube capillary had not yet solidified, the manometers showed identical readings; after that time the reading of manometer B remained constant. The measurements showed that, as in ^4He, the melting pressure of ^3He increases monotonically with temperature, rising from 40 atmospheres at 1°K to 57 atmospheres at 1·5°K. The same authors soon extended their measurements to 0·16°K (Weinstock, Abraham and Osborne 1952) using the adiabatic demagnetization of a salt for cooling the ^3He. For this process a U-tube capillary (Cu–Ni) with internal diameter 0·16 mm containing ^3He was placed in

a container filled with ferric ammonium alum. To improve the thermal contact between the capillary and the salt, copper strips were soldered to the surface of the capillary, and the container was filled with ^{4}He to a pressure of 1 atmosphere, at the temperature of liquid nitrogen. Measurements were carried out in the same order as in the previous case. The experiment showed that in the temperature interval from 0·5° to 1·5°K the melting curve is well described by the empirical formula $p = 26{\cdot}8 + 13{\cdot}1T^2$; below 0·5°K the experimental points deviated from this formula, lying on a horizontal line corresponding to a pressure $p = 29{\cdot}3$ atmospheres. The melting curve according to the observations of Weinstock, Abraham and Osborne (1952) is given in figure 9. At first the authors suggested that the

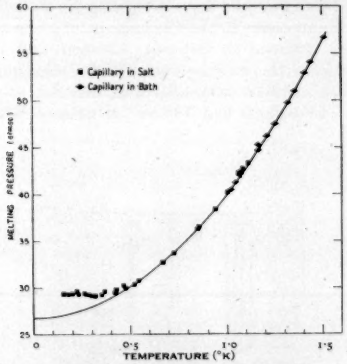


Figure 9. Melting curve of ^{3}He.

horizontal part of the curve below $T = 0{\cdot}5$°K was the result of poor thermal contact between the liquid ^{3}He and the salt, causing incomplete cooling of the liquid below 0·5°K. But it soon became apparent that the results obtained could be explained by the existence of a minimum in the melting curve, predicted by Pomeranchuk (1950) on the basis of his theory for Fermi liquids. At temperatures below this minimum, blocking of the capillary will occur at a pressure corresponding to the minimum of the melting curve. According to this, the capillary will not become blocked at the coldest point, but at points above where the temperature corresponds to the minimum of pressure. In 1957, the prediction of Pomeranchuk was indirectly verified by Walters and Fairbank (1957), who observed a rise in temperature at 0·2°K with reduction of pressure on solid ^{3}He, and at 1·2°K the effect was reversed, i.e. the reduction of pressure was followed by cooling. These observations indirectly supported Pomeranchuk's predictions, as only in this case will the heat of melting be negative.

Investigations of the melting curve at higher temperatures have been carried out by a number of authors. In the temperature region above 1·5°K (from 1·5° to 2·3°K) the solidifying curve of ^{3}He was indirectly established by Eselson and Lazarev (1954) by extrapolation from data for solidifying mixtures of ^{3}He–^{4}He.

In 1957, Sherman and Edeskuty (1957) carefully measuring the P–V–T diagram of ^{3}He, established the melting curve in the interval 1·07° to 3·10°, also using the blocked capillary method. In this temperature interval, with accuracy to within 0·2% (±0·05 atm), the melting curve is described by the equation

$$p_{at} = 23{\cdot}29 + 2{\cdot}955T + 14{\cdot}66T^2 - 1{\cdot}1871T^3. \qquad \ldots\ldots(17)$$

Mills and Grilly (1955) obtained the melting curve over a considerably wider temperature range, from 2° to 31°K, which corresponds to a change of pressure from 76 to 3500 kg cm^{-2}. Measurements were carried out also by the blocked

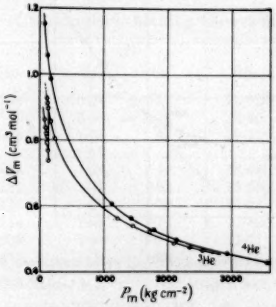


Figure 10. Measurements of volumes of ^{3}He and ^{4}He under pressure.

capillary method. The melting curve obtained by these authors was found to be a smooth continuation of the above curve, and in the interval from 2 to 31°K it is well described by the formula

$$p = a + bT^c$$

where $a = 25{\cdot}16$, $b = 20{\cdot}08$ and $c = 1{\cdot}517$. $\ldots\ldots(18)$

In 1957 these same authors (work to be published) undertook experiments to determine the change in molar volumes at melting ΔV_m, the molar volumes of the liquid V_l, and the coefficient of heat expansion of the liquid α_l along the melting curve over a wide range of pressures, from 80 to 3500 kg cm^{-2} (or from 2° to 31°K). The change in volume of solid ^{3}He under pressure was determined from the mass of gas which had to be introduced into the measuring volume during the process of melting ^{3}He to maintain constant pressure.

The authors obtained interesting and surprising results: for pressures below 141 kg cm^{-2} the curve of volume change at melting as a function of pressure appeared as two branches, an upper and a lower, displaced from each other by approximately 15%. Results of these measurements are given in figure 10. Which of these branches was obtained depended upon the initial temperature of the solid ^{3}He, and was taken as evidence of the existence of some kind of transition in it.

Careful measurements of the melting curve, at intervals of 0·1°K have been carried out by the authors in the region from 2° to 4·5°K to detect whether

there was a discontinuity in its derivative at some point; they showed that at $p = 141$ kg cm^{-2} and $T = 3{\cdot}15$°K there is in fact a small break in the slope of the curve. By further investigation it was possible to determine the part of the transition line between one solid phase and another. The transition curve was determined by careful observations of the rate of pressure increase or decrease over solid ^{3}He. At the transition point the pressure changed abruptly, giving evidence of a sudden change in molar volumes. Figure 11 shows the phase

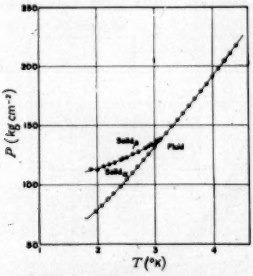


Figure 11. Melting curve and curve of phase transitions of solid ^{3}He.

diagram of solid ^{3}He, where the different modifications of the solid state are labelled α and β phases by the authors (Mills and Grilly, to be published). In table 7 are given the basic data of the α–β transition, obtained by Mills and Grilly.

Table 7. Characteristic Curve of the α–β Transition in Solid ^{3}He

P (kg cm^{-2})	T (°K)	ΔV (cm^3 mol^{-1})	ΔS (cal mol^{-1} deg^{-1})
110	1·78	0·138	0·053
120	2·35	0·142	0·063
130	2·82	0·150	0·093
141	3·15	0·157	0·138

In the light of these results it became possible to understand the existence of two branches on the curve of volume changes at melting (figure 10). When solidification of ^{3}He took place at a temperature corresponding to the α-phase, melting resulted in a volume change ΔV_m. If, on the other hand, solidification took place at a temperature lower than the transition temperature, the volume change at melting increased by a corresponding value, equal to the volume change at α–β transition. Table 8 gives the data of volume change and entropy at melting, and also the volumes and the coefficient of volumetric compressibility of the liquid phase along the melting curve of ^{3}He.

To explain the structures of the α and β phases, Shach, Grilly and Mills (to be published) applied x-ray analysis to solid ^{3}He. The authors investigated the diffraction of x-rays on solid ^{3}He contained in a beryllium capillary. They carried

out the investigation using the von Laue and Debye–Scherrer–Hull method. It was revealed that ^3He, just as ^4He, forms large single crystals considerably more easily than fine crystalline powder. By x-ray analysis it was established that the α-phase has a cubic lattice. At temperature 1·9°K and pressure 100 kg cm^{-2} the length of the cube edge is $A = 4{\cdot}01 \pm 0{\cdot}03$ Å. Calculated on the basis of these data, the density of the α phase was found to be $\rho_\alpha = 0{\cdot}154 \pm 0{\cdot}004$ g cm^{-3}, which is in good agreement with direct measurements of density by Mills and Grilly (to be published). The structure in the β phase appeared to be close-packed hexagonal crystals. At $T = 3{\cdot}3$°K and $p = 183$ kg cm^{-2} the crystal axes were equal

Table 8. Characteristic Melting Curves of ^3He and ^4He

(1)	(2)		(3)		(4)		(5)	
	^3He	^4He	^3He	^4He	^3He	^4He	^3He	^4He
80	0·854	1·150	0·94	1·37	21·92	19·74	7·5	12·0
100	0·815	1·106	0·99	1·38	20·80	19·10	7·0	10·7
141	0·737	1·036	1·06	1·43	19·72	18·18	6·5	9·5
141	0·894		1·15					
200	0·834	0·955	1·26	1·49	18·42	17·22	5·9	8·3
500	0·688	0·762	1·44	1·62	15·56	14·74	4·4	5·7
1000	0·590	0·628	1·58	1·70	13·03	13·03	3·4	4·1
1500	0·536	0·554	1·65	1·73	12·42	12·05	2·9	3·3
2000	0·494	0·506	1·69	1·75	11·71	11·38	2·6	2·8
3000	0·446	0·452	1·74	1·80	10·78	10·46	2·1	2·3

(1) P_m(kg cm^{-2}); (2) ΔV_m(cm^3 mol^{-1}); (3) ΔS_m(cal mol^{-1} deg^{-1}); (4) V_f(cm^3 mol^{-1}); (5)α_f(10^3 deg^{-1}).

to $A = 3{\cdot}46 \pm 0{\cdot}03$ Å and $C = 5{\cdot}60 \pm 0{\cdot}03$ Å which corresponds to a density $\rho_\beta = 0{\cdot}172 \pm 0{\cdot}004$ g cm^{-3}. It has thus been shown by x-ray investigations that the α and β phases possess different crystal structures.

It can readily be seen from the ^3He phase diagram that the melting curve does not reveal the slightest tendency to intersect the vapour pressure curve. In this way ^3He, like ^4He, lacks the usual triple point, in which the liquid, vapour and solid phases are in equilibrium.

§6. Density of Liquid, Gaseous and Solid ^3He

Approximate determination of the temperature dependence of the density of the liquid and its saturated vapour was among the very first experiments with liquid helium ^3He done by Sydoriak, Grilly and Hammel (1949). Direct measurements of the density of liquid and vapour were carried out with great accuracy (to 0·1% for the liquid and to 1·3% for the vapour) by Kerr (1954) in the temperature interval from 1·3° to 3·2°K as early as 1954. The equipment for measuring density (Kerr 1954) consisted of a calibrated vessel of volume 50 cm^3, containing gas at room temperature, a small Toepler pump, and a manometer for measuring pressure. Gas, the quantity of which was measured by the calibrated vessel, was lead through a Monel metal capillary, of internal diameter 0·31 mm, to a pyknometer immersed in a bath of liquid helium. In order to reduce fluctuation of the amount of gas contained in the capillary due to the temperature gradient the latter was surrounded by a copper vacuum jacket and immersed in a liquid

helium bath at the lower end and a liquid nitrogen bath at the upper end. The pyknometer was made of Pyrex glass, and had a narrow calibrated neck with a reference mark. The volume of the pyknometer up to the engraved mark ($0{\cdot}11148 \pm 0{\cdot}00006$ cm^3) had been determined by weighing it empty and filled with mercury. The pyknometer used for measurements of the density of the gas had a volume of $0{\cdot}7465 \pm 0{\cdot}0003$ cm^3. Corrections have been applied to the calculations to take account of the change of volume of the pyknometer from room temperature to helium temperature. The results of Kerr's measurements are given graphically in figure 12. The values of the density of the liquid and

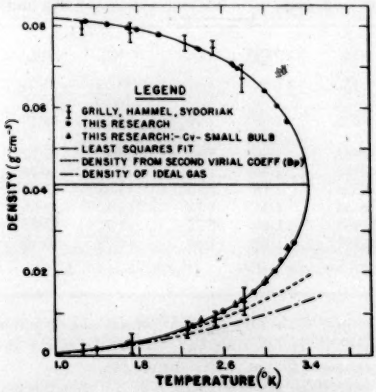


Figure 12. Curve of density of ^{3}He (*after* Kerr 1954).

vapour in the interval 1·0°–3·3°K are also given in tables 9 and 10, where the temperature is defined according to the Kistemaker (1946) scale. The table also gives the values of density calculated from the formula

$$\rho = 0{\cdot}04117 + 4{\cdot}14 \times 10^{-5}\,T \pm (0{\cdot}04117^2 - 3{\cdot}3015 \times 10^{-6}\,T^2 - 4{\cdot}1876 \times 10^{-5}\,T^3)^{1/2}.$$

The plus sign corresponds to the density of the liquid, and the minus to the density of the gas. The density of ^{3}He at the critical point, according to the measurements of Kerr, is equal to $\rho_c = 0{\cdot}0413$ g cm^{-3}. Kerr's results were subsequently verified by work of other authors. In 1957, Peshkov used an optical method to measure the temperature dependence of the density of liquid and gaseous ^{3}He in the interval 1·4°–3·2°K; this agreed with Kerr's curve. The density at the critical point, according to Peshkov's data, is $\rho_c = 0{\cdot}041 \pm 0{\cdot}001$ g cm^{-3}. The results of both Kerr's and Peshkov's measurements were verified in Ptoukha's work (1958), which determined the density of ^{3}He in the interval 1·3°–2·7°K. Values of density at absolute zero were obtained by Kerr, extrapolating his observations to low temperatures. However, it recently became known from the

experiments of Fairbank and Lee (1957) on the determination of thermal conductivity of ^{3}He that such extrapolation is unjustified because in the temperature region around 0·5°K the density curve has a maximum. The existence of a maximum was established from the sudden increase in heat transport in liquid ^{3}He in the region of 0·5°K under large heat fluxes directed downwards. This effect

Table 9. Density of Liquid ^{3}He

(1)	(2)	(3)	(4)
1·304	0·08121	0·08121	0·00
1·516	0·08051	0·08053	+0·02
1·709	0·07964	0·07971	+0·09
1·787	0·07935	0·07931	−0·05
1·998	0·07803	0·07803	0·00
2·013	0·07792	0·07792	0·00
2·378	0·07470	0·07472	+0·03
2·393	0·07444	0·07456	+0·16
2·698	0·07072	0·07054	−0·26
2·710	0·07040	0·07035	−0·07
2·748	0·06971	0·06973	+0·03
3·010	0·06434	0·06437	+0·05
3·020	0·06423	0·06412	−0·17
3·135	0·06067	0·06086	+0·31
3·223	0·05681	0·05775	+1·62

(1) T_K(°K); (2) density (obs) (g cm^{-3}); (3) density (calc) (g cm^{-3}); (4) difference (%).

Table 10. Density of Gaseous ^{3}He

(1)	(2)	(3)	(4)
1·395	0·00146	0·00152	+3·94
1·689	0·00269	0·00268	−0·37
2·305	0·00713	0·00706	−0·99
2·596	0·01059	0·01051	−0·76
2·767	0·01315	0·01317	+0·15
2·890	0·01529	0·01550	+1·35
2·991	0·01756	0·01776	+1·13
2·994	0·01774	0·01783	+0·50
3·097	0·02030	0·02058	+1·36
3·157	0·02293	0·02247	−2·04
3·160	0·02297	0·02257	−1·77
3·282	0·02727	0·02752	+0·91

(1) T_K(°K); (2) density (obs) (g cm^{-3}); (3) density (calc) (g cm^{-3}); (4) difference (%).

was absent at higher temperatures. The increase in heat transport was explained by the authors as an additional convectional heat transport, possibly when $\partial\rho/\partial T > 0$. This assumption was verified by measurements of the same authors using different apparatus (private communication), where heat was propagated upwards. With the new apparatus the effect observed above appeared at $T > 0{\cdot}5$°K.

To investigate the temperature dependence of the density below 1°K, Taylor and Kerr (1958) measured the coefficient of thermal expansion of liquid ^{3}He $\alpha = (1/V)\,dV/dT$ in the temperature interval from 0·54° to 1·01°K. Measurements were taken with a pyknometer, of volume about 3 cm suspended from a tube with small internal diameter. Using a constant mass of ^{3}He condensed in the pyknometer, the state of the meniscus of the liquid at different temperatures was determined. The coefficient of expansion had a positive value over the whole

interval investigated. The values of α for three temperatures are given below:

T (°K)	0·55	0·7	0·9
α	0·0036	0·0082	0·0164

Thus the authors had not yet reached the anomaly on the density curve. Judging from the fact that the increase in thermal conductivity in the first experiments of Fairbank and Lee (1957) was only observed at large heat fluxes, the maximum on the density curve is, in all likelihood, small. For the elucidation of the temperature dependence of density near absolute zero further experiments are necessary.

Measurements of density of liquid ^{3}He under pressure over a wide interval of temperature and pressure, have been carried out by a number of investigators.

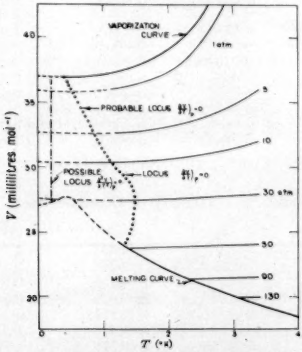


Figure 13. (v, T) diagram of ^{3}He.

Figure 13 shows the diagram of ^{3}He calculated (Hammel, Sherman, Kilpatrick and Edeskuty, to be published) from the data of Sherman and Edeskuty (1957), Taylor and Kerr (1958) and Kerr (1954).

In this diagram the solid lines represent the experimental curve, and the broken lines represent the curves obtained by extrapolation. The locus of the minima of the isobars (density maxima) are indicated by dots. Since in accordance with Nernst's theorem $(\partial V/\partial T)_p = 0$ at $T = 0$, there should be a point of inflection on each curve below the minimum, where $(\partial^2 V/\partial T^2)_p = 0$. Because of the lack of experimental data this line (broken) has been drawn arbitrarily. For the density of solid ^{3}He, as mentioned above, values at two points have been obtained by x-ray analysis:

α phase: $T = 1{\cdot}9$°K, $p = 100$ kg cm^{-2},
$\rho_\alpha = 0{\cdot}154 \pm 0{\cdot}004$ g cm^{-3};

β phase: $T = 3{\cdot}3$°K, $p = 183$ kg cm^{-2},
$\rho_\beta = 0{\cdot}172 \pm 0{\cdot}004$ g cm^{-3}.

§7. Latent Heat of Vaporization

Latent heat may be calculated from the Clausius–Clapeyron equation

$$\Delta Q_{vap} = T(V_{gas} - V_{liq})\frac{dP}{dT} \qquad \ldots\ldots(19)$$

based on experimental data of the density of the liquid and the vapour, and the temperature dependence of the saturated vapour pressure. Such calculations were made by Abraham, Osborne and Weinstock (1950), and later by Kerr (1954)

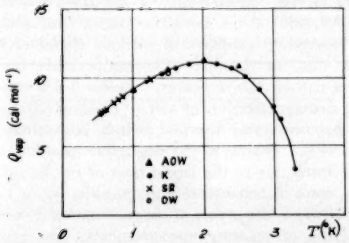


Figure 14. Latent heat of ^3He. AOW, Abraham, Osborne and Weinstock 1950; K, Kerr 1954; SR, Sydoriak and Roberts 1955; OW, Osborne and Weinstock 1958.

and Sydoriak and Roberts (1955). Direct measurements of latent heat at one point $T = 1{\cdot}5°K$ ($\Delta Q_{vap} = 10{\cdot}39 \pm 0{\cdot}02$ cal mol^{-1}) were carried out very accurately by Osborne and Weinstock (1958) using the calorimetric method. The temperature dependence of the latent heat ΔQ_{vap} according to authors previously mentioned,

Table 11. Latent Heat of ^3He

Abraham, Osborne and Weinstock (1950)		Kerr (1954)		Roberts and Sydoriak (1955)		Osborne and Weinstock (1958)	
T	ΔQ_{vap}	T	ΔQ_{vap}	T	ΔQ_{vap}	T	ΔQ_{vap}
0	(4·47)	1·0	9·10	0·54	7·36	1·5	$10{\cdot}39 \pm 0{\cdot}02$
0·5	(6·95)	1·2	9·76	0·566	7·47		
1·0	$9{\cdot}12 \pm 0{\cdot}12$	1·4	10·32	0·598	7·60		
1·5	$10{\cdot}65 \pm 0{\cdot}24$	1·6	10·71	0·629	7·72		
2·0	$11{\cdot}34 \pm 0{\cdot}41$	1·8	11·02	0·748	8·19		
2·5	$10{\cdot}81 \pm 0{\cdot}69$	2·0	11·15	0·804	8·41		
		2·2	11·05	0·872	8·66		
		2·4	10·71	1·256	9·90		
		2·6	10·11				
		2·8	9·9				
		3·0	7·88				
		3·1	7·01				
		3·2	5·91				
		3·3	3·36				

T in °K; ΔQ_{vap} in cal mol^{-1}.

is shown in figure 14. It can easily be seen that it is analogous to the temperature dependence Q_{vap} of ^4He, although approximately half in absolute value. Values of ΔQ_{vap}, obtained by different authors, are given in table 11.

§8. Specific Heat

Measurement of the specific heat of ^{3}He has been of great interest to investigators. The discovery of an anomaly in the specific heat curve would show unambiguously that there was a transition in liquid ^{3}He similar in value to that of the λ anomaly in the specific heat of ^{4}He, which is evidence for the phase transition of normal helium into a superfluid.

Specific heat in ^{3}He (96% ^{3}He, 4% ^{4}He) was first measured by de Vries and Daunt (1953) in the temperature interval 0·57–2·3°K at saturated vapour pressure. The very small quantity of ^{3}He—in all 8 cm^3 of gas at S.T.P.—which the authors had at their disposal, determined the construction of the apparatus. ^{3}He was condensed through a thin-walled capillary of internal diameter 0·25 mm into an adiabatic calorimeter, 16 mm^3 in volume. The temperature was recorded automatically, by means of a carbon thermometer. Below 1·3°K the calorimeter was cooled by the adiabatic demagnetization of salt. Thermal contact with the bath was established by a superconducting thermal switch, controlled by the magnetic field (Heer and Daunt 1949). Owing to the negligible quantity of ^{3}He contained in the calorimeter, corrections due to the latent heat of the liquid, specific heat of the vapour and so forth, were of considerable magnitude, up to 17%. According to the assessment of the authors, the errors of measurements do not exceed $\pm$ 10%. Despite the wide scattering of separate measurements, the results quite clearly showed the smooth decrease of specific heat in the temperature interval investigated from 3·6 to 0·9 cal deg^{-1} mol^{-1}, and showed the complete absence of any kind of anomaly. In these measurements, the temperature region below 1·0°K attracted attention, where the specific heat had a constant value equal to 0·9 cal deg^{-1} mol^{-1}. Soon Sydoriak and Roberts (1954, 1955) and simultaneously Abraham, Osborne and Weinstock (1955, Osborne, Abraham and Weinstock 1954) in their own measurements of specific heat, extended the temperature interval to 0·37° and 0·23°K respectively. Measurements of Roberts and Sydoriak (1955) were carried out in the region from 0·37° to 2·36°K with the apparatus they had used earlier for the determination of saturated vapour pressure (figure 6). For determining specific heat the authors modified the usual calorimetric method. ^{3}He with a ^{4}He content of less than 0·1%, was condensed in a copper sphere of internal diameter 10 mm, suspended on a thin-walled pipe (Cu–Ni) within a vacuum. The sphere was filled up to 50% of its volume with paramagnetic salt. The calorimeter was cooled to a temperature of 0·5°K by pumping ^{3}He vapour; lower temperatures were obtained by the adiabatic demagnetization of salt. The temperature of the liquid was determined from the magnetic susceptibility of the salt.

The process of measurement consisted in the determination of the heating rate of the liquid $\dot{T}$ by a known heat input $\dot{Q}$ depending on the quantity of ^{3}He contained in the calorimeter. It is easy to see that the slope $\dot{Q}/\dot{T}$ as a function of N the number of moles of ^{3}He (after taking into account the corrections in the latent heat of the liquid, and the specific heat of the vapour in the calorimeter) gives directly the specific heat of the liquid in cal deg^{-1} mol^{-1}. The specific heat of the empty calorimeter is automatically excluded in this method of measurement. The method, however, is of little use for measurements at high temperatures, where the corrections become very large. In the temperature interval 0·5°–1·7°K the accuracy of measurements according to the authors' estimate is to 1·5–2%.

Abraham, Osborne and Weinstock took measurements in the temperature interval 0·23°–2°K. The adiabatic calorimeter had volume 1·87 cm³. ^3He was cooled by adiabatic demagnetization of salt, which was in contact with the liquid. The value of specific heat in the work of Abraham, Osborne and Weinstock (1955) was determined with an accuracy to within ±5%.

The results of the work of all three groups of authors are in agreement within the limits of experimental error, with the exception of those of de Vries and Daunt (1954) in the region below 1°K. According to more accurate measurements (Roberts and Sydoriak 1955, Abraham, Osborne and Weinstock 1955) the specific heat of ^3He in the region below 1°K decreases from 1·0 cal deg^{-1} mol^{-1} to 0·7 cal deg^{-1} mol^{-1} at 0·25°–0·3°K. Results of the measurements of Sydoriak and Roberts (1955), Abraham, Osborne and Weinstock (1955) are given in figure 15.

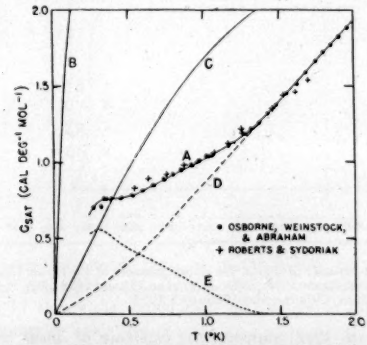


Figure 15. Specific heat of liquid ^3He under saturated vapour pressure. A, experimental results; B, C, C_v of an ideal Fermi–Dirac gas with corresponding temperature degeneracy 0·45° and 4·98°K; D, calculated curve of non-spin thermal capacity; E, calculated curve of spin thermal capacity (curve A − curve D).

In 1958 the work of Brewer, Sreedhar, Kramers and Daunt appeared, in which measurements of specific heat of ^3He were carried out under pressures of from 12 to 15 cm Hg, in the temperature interval 0·7°–0·085°K.

^3He was condensed in an adiabatic calorimeter of volume 0·5 cm³. The calorimeter was cooled by adiabatic demagnetization of salt, the latter being in contact with the liquid ^3He. Thermal contact with the bath was established by means of a superconductive lead switch, controlled by the magnetic field (Heer and Daunt 1949). A manganin wire of resistance 100 Ω served as an electric heater. The temperature was measured by the magnetic susceptibility of the cerium magnesium nitrate. This salt was selected because of its close adherence to Curie's law in the interval investigated, and also because it has a negligible specific heat.

The results of the measurements of Brewer, Sreedhar, Kramers and Daunt (1958) are given in figure 16 where, for comparison, data of Abraham, Osborne

and Weinstock (1955) and Roberts and Sydoriak (1955) are also given. It can be seen from the diagram that the results of Brewer *et al.* (1958) appear as a smooth continuation of the preceding curve, shown in figure 15 (the difference in specific heats results from the difference in pressure over the liquid, which is here very small). The linear extrapolation of results towards 0°K in this diagram is typical.

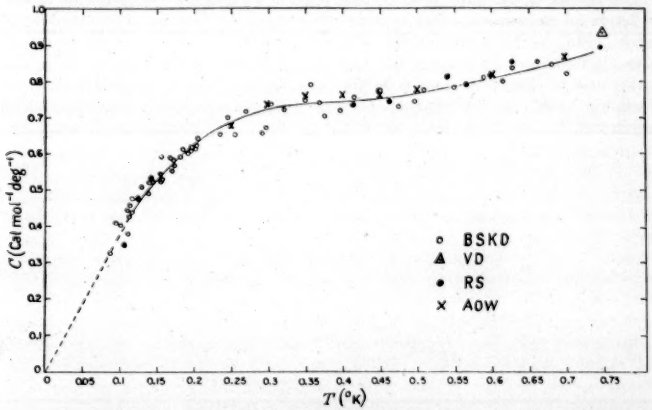


Figure 16. Molar thermal capacity of liquid ^{3}He under pressure of 12–15 cm Hg. BSKD, Brewer, Sreedhar, Kramers and Daunt 1958 ; VD, de Vries and Daunt 1954 ; RS, Roberts and Sydoriak 1955 ; AOW, Abraham, Osborne and Weinstock 1955.

The work (Brewer *et al.* 1958) supports the existence of linear terms in specific heat, predicted for Fermi fluids by Pomeranchuk (1950) and obtained on the basis of Landau's theory, by Khalatnikov and Abrikosov (1957). With this, all doubts of the possibility of ^{3}He transition into the superfluid state, above 0·1°K, disappear, since in this region the specific heat curve does not reveal any anomaly.

§9. Entropy

The data on specific heat (specific heat below 0·08°K was assumed to be proportional to T) enabled the authors to calculate the entropy of liquid ^{3}He in the interval 0–0·7°K (Brewer *et al.* 1958). Values of entropy at higher temperatures were calculated by Roberts and Sydoriak (1955), Abraham, Osborne and Weinstock (1955) and de Vries and Daunt (1954) based on the specific heat data, and on existing experimental data on vapour pressure and latent heat in ^{3}He. In the temperature interval 0·085–2·5°K the entropy of the liquid, according to calculations of these authors, increases smoothly from 0·3 cal deg^{-1} mol^{-1} to 9·5 cal deg^{-1} mol^{-1}. Below 0·1°K the curve has a linear section with a slope 3·75 cal mol^{-1} deg^{-2}.

The temperature dependence of the entropy of liquid and solid ^{3}He was considered theoretically by Pomeranchuk (1950). According to his theory, the

entropy of the liquid at temperatures of the order of 1°K should tend to zero as $T \to 0$, considerably more rapidly than does the entropy of the solid state. This is due to the fact that in the liquid, the exchange interaction forces lead to the alignment of spin orientations, parallel or antiparallel, at temperatures below 1°K, whilst in the solid, owing to the small amplitude of the zero point oscillations compared with interatomic distances, alignment takes place at considerably lower temperatures (10^{-6}–10^{-7} deg K) when kT becomes of the order of the magnetic energy of interaction of two neighbouring atoms. Because of this the entropy of solid ^3He in the temperature interval from 10^{-7} to 1°K is basically determined by the random distribution of nuclear spin orientations and should be equal to $R \ln 2 = 1{\cdot}377$ cal deg^{-1} mol^{-1} (the part of the entropy corresponding to thermal

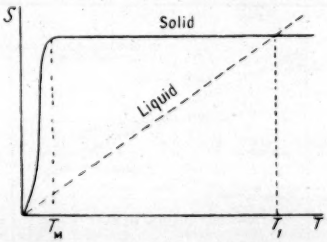


Figure 17. Entropy of solid and liquid ^3He according to Pomeranchuk (1950).

excitations is practically zero), then, as in the liquid, entropy falls rapidly with further decrease in temperature corresponding to increase in the degree of alignment. Figure 17 shows the dependence of entropy of liquid and solid ^3He, as predicted by Pomeranchuk's theory.

In consequence of this temperature dependence of entropy in liquid and solid ^3He a minimum appears on the melting curve, resulting from the Clausius–Clapeyron equation :

$$\frac{dp_m}{dT} = \frac{S_{liq} - S_{sol}}{V_{liq} - V_{sol}} < 0 \quad (T < T_{min}),$$

$$\frac{dp_m}{dT} = \frac{S_{liq} - S_{sol}}{V_{liq} - V_{sol}} > 0 \quad (T > T_{min}). \qquad \ldots\ldots(20)$$

As can readily be seen from the relations given in (20), the position of the minimum on the melting curve corresponds to the temperature at which the entropies of the liquid and solid phases in equilibrium with each other become equal. From the relations (20) it also follows that the sign of the latent heat of melting will change during transition through the minimum

$$Q_m = T(S_{liq} - S_{sol}) < 0 \quad (T < T_{min}),$$

$$Q_m = T(S_{liq} - S_{sol}) > 0 \quad (T > T_{min}). \qquad \ldots\ldots(21)$$

It follows from equation (21) that at temperatures below T_{min} the solidification of ^3He must be accompanied by heat absorption in the solid and, consequently, the adiabatic solidification by compression of ^3He should lead to a reduction of

temperature: this was called 'Pomeranchuk's effect'. The behaviour of entropy of solid ^{3}He at low temperatures was also considered by Primakoff (1957). According to Primakoff's theory, the temperature at which nuclear spins in solid ^{3}He align themselves antiparallel is lower the higher the pressure: thus under pressures of 30 atmospheres spin orientation begins at 0·5°K, while at 10^3 atmospheres orientation takes place at $T = 10^{-3}$ deg K. Accordingly the entropy of solid ^{3}He increases with pressure, and with adiabatic compression of ^{3}He cooling should result. However, in contrast to Pomeranchuk, who predicted very intense cooling down to $T = 10^{-6}$–10^{-7} deg K, Primakoff predicts that cooling in the same conditions should be considerably less.

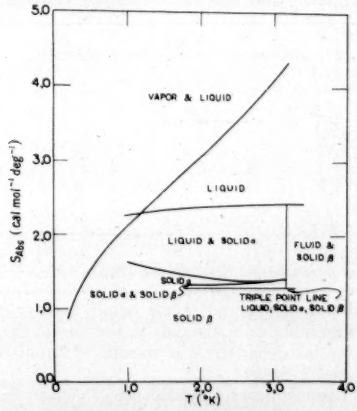


Figure 18. Entropy diagram of ^{3}He.

From what has been said above, the attempt to explain the nature of the ^{3}He entropy diagram becomes understandable. Unfortunately, there are as yet no direct measurements either of the specific heat of liquid under pressure or of the specific heat of solid ^{3}He which would enable direct verification of the predictions of the theory. However, on the basis of existing experimental data we can make preliminary calculations of the entropy of solid ^{3}He and liquid ^{3}He under pressure. Figure 18 shows the entropy diagram of ^{3}He as calculated by Hammel, Sherman, Kilpatrick and Edeskuty (1958) based on all the present experimental data. In this diagram, the entropy of the liquid at $T = 1{\cdot}5$°K and saturated vapour pressure $S = 2{\cdot}614 \pm 0{\cdot}03$ cal deg^{-1} mol^{-1} was calculated from the values of the latent heat of vaporization of ^{3}He, measured for this purpose with great accuracy by the calorimetric method by Osborne and Weinstock (1958), and of the entropy of the vapour. The calculations are based on the formula

$$S_{\text{liq}} = S_{\text{vap}} - \frac{\Delta Q_{\text{vap}}}{T}. \qquad \ldots\ldots(22)$$

In all the remaining temperature intervals the entropy of the liquid under saturated vapour pressure was determined from the specific heat data of Roberts and Sydoriak (1955) and Brewer, Sreedhar, Kramers and Daunt (1958). Changes of the entropy of the liquid with pressure have been calculated from the formula

$$\Delta S_{\text{comp}} = \int_{p_{\text{sat}}}^{p} \left(\frac{\partial V}{\partial T}\right)_p dp \qquad \ldots\ldots(23)$$

where the value $(\partial V/\partial T)_p$, and the function p have been taken from the work of Sherman and Edeskuty (1957). In this way separate isobars were obtained (not shown in figure 18) within the entire range of existence of the liquid, from saturated

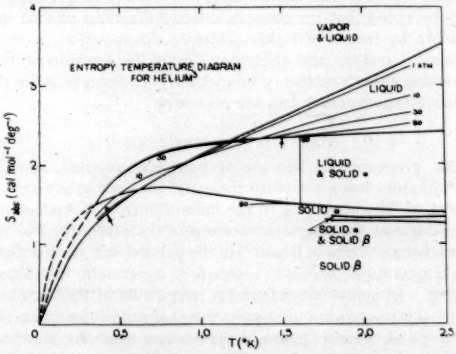


Figure 19. Entropy diagram of ^3He.

vapour pressure up to the pressure of solidification. The entropy of the liquid in equilibrium with the solid phase was calculated from the results of Weinstock, Abraham and Osborne (1955), Sherman and Edeskuty (1957) and Grilly and Mills (1958). The abrupt discontinuity of entropy during transition from the liquid to the solid state, in the region of low temperatures (from 1·9°K and above) was calculated from the data of Grilly and Mills (1958) who measured dp_m/dT and the changes in molar volume at melting. At the lowest temperatures, owing to the lack of experimental data, the curves of changes in molar volume and of entropy were obtained by extrapolation. The findings of Grilly and Mills (1958) also enabled the prediction of the abrupt change in entropy of solid ^3He during transition from α to β phase.

Consideration of the entropy diagram of ^3He shows that with decreasing temperature the entropy along the solidifying curve falls, but along the melting curve in the region of comparatively high temperatures it remains constant and even increases slightly. By extrapolating the curves in the direction of low temperatures, the authors have been able to show that the curves intersect at $T = 0{\cdot}35$°K, where according to the thermo-dynamical relationships (20) there should be a minimum on the melting curve. Figure 19 shows the entropy

diagram of ^{3}He derived from the calculations of Hammel, Sherman, Kilpatrick and Edeskuty (1958) consisting of a series of isobars, and completed by means of extrapolation of the curves in the low-temperature region. Extrapolation was carried out on the basis of known experimental data, and theoretical predictions (Primakoff 1957, Brueckner and Hammel 1957). In this diagram the section of the entropy curve of solid ^{3}He below 0·3°K may give rise to some doubt. The curve is based on the experimental data of Fairbank and Walters (1957 b) on the magnetic susceptibility of ^{3}He and Primakoff's theory (1957). There is, however, some reason to suppose that the measurements of Fairbank and Walters, which indicate that spin alignment in solid ^{3}He begins to appear at 0·3°K, are not very reliable. We will consider the experiments of Walters and Fairbank in detail below. It may be noted that the present entropy diagram, plotted in the region of low temperature by means of rather arbitrary extrapolation, is definitely only an approximation to reality, and therefore cannot be considered to give good grounds for choosing either one theory or another. In order to solve the problem, new and conclusive experimental data are necessary.

§ 10. Magnetic Properties of ^{3}He

The magnetic properties of ^{3}He are of particular interest, since, as distinct from ^{4}He, the ^{3}He atom has a non-zero magnetic moment and a spin of $\frac{1}{2}$. The magnetic moment of ^{3}He, according to the measurements of Anderson and Novik (1948) is $0{\cdot}763 \pm 0{\cdot}007$ of the magnetic moment of the proton or $1{\cdot}07 \times 10^{-23}$ c.g.s. Because of the exchange effects in liquid ^{3}He the parallel and antiparallel spin alignment begins at temperatures below 1°; this is in agreement with Pomeranchuk's predictions (1950). In solid ^{3}He, where the amplitude of the zero point oscillations is less than the interatomic distance, the exchange interaction is practically absent, and, because of the interaction of the nuclear magnetic moments, the spin alignment should only begin at temperatures of the order of 10^{-6} degree. In contrast to Pomeranchuk, Primakoff (1957) proposed that in solid ^{3}He at temperatures below 0·6°K the atoms already begin antiparallel spin alignment. It should be pointed out that a considerable number of theoretical papers have been devoted to the problem of nuclear spin alignment. Their criterion for correctness can perhaps be seen in the investigation of the magnetic properties of ^{3}He.

The first experiments on the determination of the magnetic properties of liquid ^{3}He were carried out by Sydoriak and Hammel (1949), and Hammel, Laquer, Sydoriak and McGee (1952). Measurements were carried out in two ways. In the first method, they observed the displacement of the frequency of an oscillator working at 1·47 Mc/s, when ^{3}He was condensed within its cell over a temperature range from 2·08°K to 1·27°K. With accuracy to 5 c/s, there was no change, which shows that the volume magnetic susceptibility of liquid ^{3}He was less than 5×10^{-6}.

This method would not give positive results if ^{3}He were ferromagnetic and was orientated by the earth's magnetic field, for a variable field would not have time to reorientate the domains, because of the large relaxation time. To check this possibility, the authors carried out experiments at a frequency of 60 c/s. They put two matched coils connected in opposition in an oscillating field of 0·65 oersted and measured the signal when ^{3}He was condensed in one of them, at $T = 0{\cdot}904$°K;

in another case condensation took place at $T = 2{\cdot}51°K$, whereupon, in the course of 45 minutes, the temperature fell to $1{\cdot}35°K$. No effect was observed in either case, within the limits of experimental error ($\sim 5\%$). Although these experiments were not of sufficient sensitivity to measure the magnetic permeability of liquid ^3He, nevertheless they enabled it to be shown that ^3He does not become ferromagnetic down to $T = 0{\cdot}9°K$.

A more sensitive apparatus, using the paramagnetic resonance method discovered by Zavoisky (1945), was made by Fairbank and co-workers (Fairbank and Walters 1957a, Fairbank, Ard and Walters 1954, Fairbank *et al.* 1953). The most complete account of the experimental work is contained in the reports of the

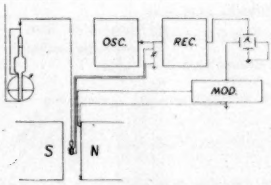


Figure 20. Simplified system for measuring paramagnetic resonance in ^3He.

1957 conference at Ohio State University (Fairbank and Walters 1957a). The block diagram of Fairbank's apparatus is shown in figure 20. The specimen was placed inside a coil which formed part of a tuned resonance circuit. Perpendicular to the axis of the coil there was a magnetic field of 9000 gauss. A.C. was fed to the coil by a generator working at a stabilized frequency of 30 Mc/s. The resonance circuit was fed through a high impedance network, thus enabling the input amplitude to be kept constant. Under these conditions the signal taken from the coil is proportional to the Q factor of the resonance circuit. By means of a modulator, the signal of which was fed simultaneously to the display of an oscilloscope and to an additional coil at the magnet poles, the magnetic field could be varied slightly at around 9000 gauss. At paramagnetic resonance there was increased absorption of high frequency oscillations in the specimen, and a corresponding reduction of the Q factor of the circuit. The signal output was amplified, detected and fed to the oscilloscope, where at resonance a clearly visible maximum appeared. The area under the resonance curve on the oscilloscope is proportional to the magnetic susceptibility of the specimen. Since the natural width of the resonance line of ^3He is considerably less than the width due to the inhomogeneity of the magnetic field of the specimen, the form of the resonance curve on the oscilloscope remains unchanged throughout the experiment, and therefore the relative change of magnetic susceptibility of the specimen was directly proportional to the amplitude of the maximum. To measure the susceptibility of liquid ^3He above $1{\cdot}2°K$ (Fairbank *et al.* 1954), and that of the gas at $4{\cdot}2°K$, a small glass bulb was connected (figure 20) by a glass capillary to a Toepler pump. A copper coil was wound directly on to the bulb.

To measure the susceptibility at temperatures below 1°K (Fairbank *et al.* 1954) the apparatus shown in figure 21 was used. Liquid ^{3}He was placed inside a glass tube of volume 15 mm^3 situated in a copper cavity. The coil for nuclear resonance was inside the glass tube. Low temperatures were obtained by demagnetization of 25 grammes of chrome potassium alum pressed, as can be seen from the diagram, around the copper cavity at an appreciable distance from the measuring vessel. The ^{3}He used in these experiments was concentrated by fractional distillation to a purity higher than 99%.

In this apparatus the volume magnetic susceptibility was measured directly in arbitrary units. From these data and those of Kerr (1954) for the density of

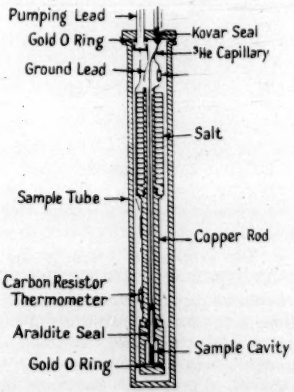


Figure 21. Apparatus for measuring susceptibility of ^{3}He at temperatures below 1°K.

liquid ^{3}He, the relative molar magnetic susceptibility was calculated. Below 1·2°K the density of liquid ^{3}He was taken as constant, because an estimate (Walters and Fairbank 1956 b) of the variations of density of liquid ^{3}He with change of volume of the compressed vapour as the specimen warmed up from 0·2° to 1·2°K showed that for all pressures and temperatures from 0·2° to 1·2°K the density of the liquid ^{3}He changes by less than 1%.

Results of the measurements are shown in figure 22. The lower curve corresponds to measurements of magnetic susceptibility of liquid ^{3}He at saturated vapour pressures. It should be pointed out that, within the limits of measurement, the lower curve coincides with the theoretical curve of an ideal Fermi–Dirac gas, having a degeneracy temperature of $T_0 = 0{\cdot}45$°K. Measurements for solid ^{3}He are shown in the same diagram; however, at low temperatures the results correspond more closely to liquid than to solid helium. In the apparatus described above the ^{3}He may melt, since its melting pressure at $T = 0{\cdot}3$° increases with decrease in temperature. Furthermore, the authors themselves point out

that their results below 0·4° sometimes give positive and not negative deviations from Curie's law of the magnetic susceptibility of solid ^{3}He (Fairbank and Adams 1958). The existence of these two values of susceptibility may be explained also by the fact that in this temperature region the existence of one of the two observed phases, α or β, of solid ^{3}He, is permissible (Mills and Grilly, to be published).

Besides the measurements quoted, Walters and Fairbank (1956), using apparatus similar to that in figure 21, determined the pressure dependence of the volume magnetic susceptibility and of the density of liquid ^{3}He at $T = 1{\cdot}2°$K. The authors assumed that under these conditions the magnetic susceptibility was proportional to density, which, within the limits of measurement (5%), was confirmed by measuring the quantity of ^{3}He which was condensed in the small bulb.

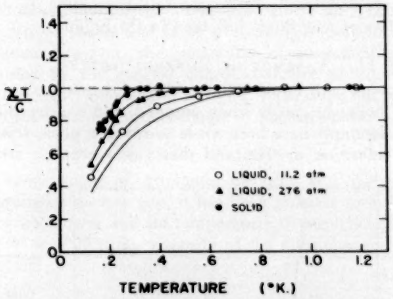


Figure 22. Dependence of $\chi T/C$ on temperature for liquid (at 11·2 and 27·6 atm) and solid ^{3}He (χ = molar susceptibility, C = Curie constant).

This was also confirmed by the density measurements of ^{3}He made by Sherman and Edeskuty (1957).

Fairbank and Walters (1957 a) also measured the relaxation time for the establishment of equilibrium of the magnetic state of liquid ^{3}He, after a considerable drop in the a.c. field amplitude, originally maintained at a level which caused saturation of the nuclear resonance. The relaxation time T_1 was the time the signal needed to increase from zero to 1/e of its final magnitude. At saturated vapour pressure in liquid ^{3}He, T_1 was not reproducible from experiment to experiment, and varied from 20 to 200 sec. This latter point led the authors to think that T_1 is determined by the influence of the wall. However, an investigation of the pressure dependence of T_1 taken over one experiment showed that it does not increase with pressure, as would have been the case with wall influence, but decreases instead.

The decrease in T_1 at first sight conforms with the theory of Bloemberger, Purcell and Pound (1948) who, taking into account the Brownian movement, obtained $T_1 \sim Da/\rho$, where D is the self-diffusion coefficient, ρ the density of ^{3}He and a the effective radius of the ^{3}He atom. Fairbank and Walters, using the experimental data for density (Kerr 1954), and for diffusion (Garwin, Kan and

Reich 1955), and assuming $a \sim 1$ Å, obtained $T_1 \sim 200$ sec. However, the irreproducibility of experiments makes the latter conclusion unconvincing. It should be conceded that the relaxation is determined by the impurities, the quantity of which is proportional to the density of liquid ^{3}He. Their measurements also show a sharp drop in the relaxation time T_1 in the solid phase down to 2 sec. Garwin and Reich (1958) reported at the Low Temperature Conference in Leiden (June 1958) that they had made measurements of diffusion D and relaxation time T_2, corresponding to the interaction of magnetic nuclei of ^{3}He. Using the spin echo technique, they found that D falls slightly from 3° to 1·8°K, and by extrapolating from 1·1° to 0° they obtained the final value $D = 2 \times 10^{-5}$ cm^2 sec^{-1}. D varies with molar volume V in the liquid as V^3, over a factor of 6 in D. Simultaneous measurements of D and T_2 in co-existing fluid and solid ^{3}He showed that for the same pressures and temperatures the relaxation time T_2 and D in solid helium are three orders of magnitude less than in the liquid.

§ 11. Absence of Superfluidity

The first experiments with liquid ^{3}He established that it is a colourless, transparent liquid, which boils vigorously when pumped, and is externally no different from He I. Several attempts have been made to reveal a phase transition in ^{3}He, analogous to the λ transition in ^{4}He, and these gave negative results. Among

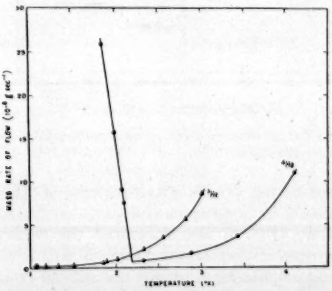


Figure 23. The velocity of transport of a mass of liquid ^{3}He and liquid ^{4}He through a 7×10^{-3} cm annulus, as a function of temperature.

these attempts are those of Abraham, Osborne and Weinstock (1949), who investigated the flow of liquid ^{3}He through a gap approximately 1 μ wide in the temperature interval 3·02°–1·05°K. The gap was formed by shrinking a Pyrex glass capillary around a platinum wire. Because of the difference in expansion coefficients, this construction had a super-leak at low temperatures, the magnitude of which was determined by the velocity of gas flow at $T = 4{\cdot}2$°K. The capillary connected two elbows of a U-tube, in one of which was condensed the liquid to be investigated. Flow of the liquid was determined by observing the rate of pressure increase in an evacuated vessel connected to the other end of the capillary. As distinct from ^{4}He, whose rate of flow through the small gap increased sharply at

the λ transition, for liquid ^3He a uniform drop in the velocity with decrease in temperature was observed, and this is evidence of the lack of superfluidity in the interval 1·05°–3·02°K. Results of these measurements are shown in figure 23. By further investigations of the properties of ^3He it was shown that ^3He does not reveal superfluidity below 1°K. Thus from measurements of viscosity and thermal conductivity of ^3He, which will be discussed below, it follows that ^3He behaves like an ordinary liquid right down to the temperatures 0·2°–0·3°K. Experiments determining the λ line and the solution separation diagram of ^3He–^4He lead to the same results (Walters and Fairbank 1956a, Roberts and Sydoriak 1958, Zinov'eva and Peshkov 1959). These experiments show that solutions with concentrations higher than 82% (Zinov'eva and Peshkov 1959) remain normal right down to the lowest temperatures at which investigations have been carried out (0·4°K). Extrapolation of the λ transition curve for mixtures to the region of higher concentrations shows that in all probability superfluidity in pure ^3He does not occur. Since, however, the possibility of superfluidity in ^3He is not in principle excluded by the general quantum theory of liquids, to reach a final conclusion further experiments at lower temperatures are needed, and in particular, experiments in the region of ordered state of nuclear spins.

§ 12. Viscosity

From the observations of Abraham, Osborne and Weinstock (1949) for the flow of ^3He through a narrow gap, it has been possible to estimate the magnitude of the viscosity coefficient and its temperature dependence. As the temperature drops from 2·79° to 1·04°K the viscosity of the fluid changed from 22 to 30·4 μP.

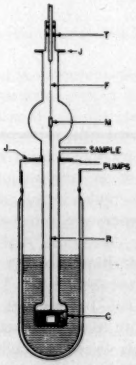


Figure 24. Apparatus for determination of viscosity of ^3He (Taylor and Dash 1957).

Quantitative measurements of the viscosity were made considerably later. In 1957 Taylor and Dash measured the viscosity of ^3He by the method of torsional oscillations of a disc, in the temperature interval 1·3°–2·15°K and at $T = 3·34$°K.

Their apparatus is shown in figure 24. A duralumin disc, of diameter 12·72 mm and thickness 1·34 mm was securely fixed to a Pyrex glass rod R, 76 cm long and 0·8 mm in diameter, which was suspended on a platinum thread F in a glass chamber C filled with ^{3}He. The chamber was 20 mm high, and 25 mm in diameter. Between the disc and the bottom of the chamber was a distance of 5 mm. The temperature of the liquid was determined from the vapour pressure of ^{3}He. The decrement of damping of the disc was measured by the mirror and scale method. To eliminate the influence of mechanical vibrations, the cryostat was fixed securely to a massive steel frame resting on a layer of cork. The effect of electric fields created by the static charges on the walls of the chamber and the

Table 12. Viscosity of Liquid ^{3}He

Taylor and Dash (1957)		Zinov'eva (1958)			
T (°K)	$\eta \times 10^5$ (poise)	T (°K)	$\eta \times 10^5$ (poise)	T (°K)	$\eta \times 10^5$ (poise)
1·299	2·640	0·35	4·8	1·4	2·02
1·396	2·508	0·40	4·00	1·6	1·96
1·451	2·570	0·45	3·64	1·8	1·92
1·500	2·671	0·50	3·40	2·0	1·87
1·599	2·493	0·55	3·22	2·2	1·83
1·601	2·515	0·5	3·07	2·4	1·78
1·697	2·328	0·7	2·82	2·6	1·74
1·797	2·282	0·8	2·63	2·8	1·70
1·893	2·292	0·9	2·46	3·0	1·66
1·951	2·295	1·0	2·33	3·2	1·61
2·000	2·360	1·2	2·14		
2·152	2·333				
2·154	2·246				
3·338	1·879				
Saturated vapour					
3·35	1·860				

Dewar was eliminated by covering the internal walls of the chamber and the surface of the rod R with a semi-transparent coating of silver, which ensured electrical contact between the disc and the chamber. Owing to the formation of bubbles and convection currents in the liquid, the reproducibility of results was not very good. Within the temperature interval 1·3°–2·2°K the standard deviation of the results was $\pm 2{\cdot}5\%$. Above 2·2°K the deviation increased to $\pm 10\%$, and the authors did not consider it possible to include these data in the table, except for the value $T = 3{\cdot}35$°K (see table 12). The measurements of the viscosity coefficient by Taylor and Dash (1957) agree with those of Abraham, Osborne and Weinstock (1949), both in absolute magnitude and in temperature dependence.

In 1958 measurements of the viscosity coefficient over a wider temperature interval from 0·35°–3·2°K were made by Zinov'eva (1958), by the capillary viscometer method. Figure 25 shows one of the instruments used for measurements of the viscosity coefficient at low temperatures. It consisted of a Dewar i of volume about 3 cm^3 and diameter 8 mm from which ^{3}He vapour was pumped out. The liquid flowed through a thin-walled capillary a, to the end of which

a cylindrical reservoir b was cemented. To shield the capillary from impurities, filters e and d were placed at each end. These filters had a large surface area (diameter 5 mm) and were made of the finest cotton wool, which would trap particles of the order of 0·1 μ. The fluid temperature of the liquid was determined by means of a phosphor-bronze resistance thermometer R_1. The reservoir b together with the capillary was suspended on the thread g and was raised by a magnet (not shown in figure 25). Measurements were made in the following way. After filling the reservoir b with liquid, the capillary was raised sufficiently for its lower end to be clear of the liquid. With a cathetometer and a stop watch, measurements were made of the change with time of the level in the reservoir.

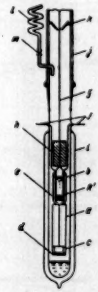


Figure 25. Capillary viscometer for ^3He (Zinov'eva 1958).

Except at the lowest temperatures, measurements were made with the pump switched off. So that the temperature would not change materially during the course of a measurement which lasted 1–2 minutes, a pellet of paramagnetic salt h was fixed on to the reservoir, which was cooled by ^3He vapour. Because of its large thermal capacity the salt acted as a buffer which reduced the rate of heating of the liquid. Measurements were made on two capillaries of diameter 100·2 μ and 74·5 μ. Absolute values of viscosity coefficient were determined in this work, with an accuracy to $\pm 5\%$ in the region above 1°K and to $\pm 10\%$ at the very lowest temperatures. The results of Zinov'eva's measurements are shown in figure 26, together with data of Abraham, Osborne and Weinstock (1949) and Taylor and Dash (1957). Values of the viscosity coefficient found by various authors are given in table 12. Divergence between the results of Taylor and Dash (1957) and Zinov'eva (1958) is of the order of 20% and may in all probability be explained by systematic error in the determination of the constants of the instruments. As may be seen from figure 26, the viscosity of ^3He changes little with temperature between the critical point and 1°K, but suddenly increases sharply with drop in temperature below 1°K, reaching the value $4{\cdot}8 \times 10^{-5}$ poise at $T = 0{\cdot}35$°K. In 1950 Pomeranchuk (1950) from qualitative considerations predicted a dependence of the form $1/T^2$ for the viscosity of liquid ^3He. Later Abrikosov and Khalatnikov

(1956, 1958) from a detailed consideration of the kinetic coefficients of liquid ^{3}He, and on the basis of the theory of a Fermi liquid suggested by Landau (1956), showed that the viscosity must have a dependence of the form $1/T^2$ below 0·1°–0·05°K. An analysis of the curve of figure 26 shows that below 1°K the viscosity changes as $1/T^{1/2}$ and it is only in the region 0·4°–0·5°K that the dependence becomes steeper ; however, it is still weaker than $1/T^2$, which does not contradict the estimate given by Abrikosov and Khalatnikov (1956).

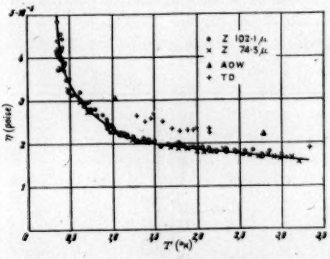


Figure 26. Viscosity of liquid ^{3}He : Z, Zinov'eva 1958 for capillary diameters as shown ; AOW, Abraham, Osborne and Weinstock 1949 ; TD, Taylor and Dash 1957.

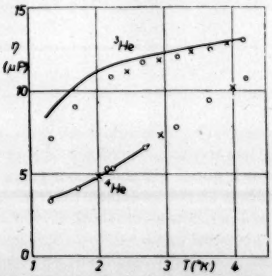


Figure 27. Temperature dependence of the viscosity of gaseous ^{3}He and ^{4}He. Circles and crosses, experimental results obtained with decrease and increase of temperature, respectively. Solid lines, theoretical curves of de Boer and Cohen.

The viscosity of gaseous ^{3}He was determined considerably earlier than that of liquid ^{3}He. Becker and Misenta (1954) made measurements over the temperature range 1·3°–4·2°K by the oscillating cylinder method. The accuracy of these measurements was about to 1%. Figure 27 shows the viscosity coefficients of gaseous ^{3}He and ^{4}He, as determined by Becker and Misenta (1954). The results obtained by decreasing and increasing the temperature are denoted by circles and crosses respectively. The continuous lines represent the theoretical curves calculated by de Boer and Cohen (1951).

§13. Thermal Conductivity

In 1957, Lee, Donnelly and Fairbank (1957, Fairbank and Lee 1957) measured the thermal conductivity of liquid ³He (under pressure of about 2/3 atmosphere) over the temperature range 0·2°–2·7°K. They used apparatus (figure 28) similar to that of Fairbank and Wilks (1955) to measure the thermal conductivity of ⁴He below 1°K. ³He was condensed in a thin-walled cupro-nickel capillary A (internal diameter 1·39 mm, length 2 cm) surrounded by a vacuum jacket J. On the upper welded end of the capillary was an electric heater H ; the lower end of the capillary was connected through copper C with a pellet P of chrome potassium alum, which was cooled by adiabatic demagnetization. The salt and the capillary with its

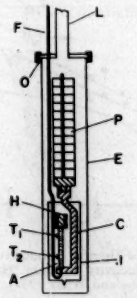


Figure 28. Apparatus for determination of thermal conductivity of ³He (Lee, Donnelly and Fairbank 1957).

vacuum jacket were isolated from the bath of liquid helium by an hermetically sealed chamber. When heat was propagated downwards along the capillary, a temperature difference appeared in the column of liquid, which was measured by two carbon thermometers T_1 and T_2, placed on the surface of the capillary. A correction for the thermal conductivity of the tube walls was made by measuring the thermal conductivity of the empty capillary. (This correction was rather large, because the thermal conductivity of the tube was greater than that of the column of liquid.) The results of Fairbank and Lee are given in figure 29, black circles representing ³He and open circles ⁴He. It can be seen that the thermal conductivity of ³He in the temperature interval 0·2°–2·7°K, increases smoothly from $0{\cdot}7 \times 10^{-4}$ to $1{\cdot}9 \times 10^{-4}$ watt cm^{-1} deg^{-1}. The expected temperature dependence of the thermal conductivity of ³He is of the form $1/T$; this follows from Pomeranchuk's theory (1950) and also from that of Abrikosov and Khalatnikov (1956), but this evidently occurs below 0·2°K.

Results of Lee, Donnelly and Fairbank were simultaneously confirmed by Challis and Wilks (1957), who measured the thermal conductivity of liquid ³He (95% ³He, 5% ⁴He) in the temperature interval 1·3°–3·0°K. The authors had only small quantities of ³He and this determined the construction of their apparatus, which did not enable them to eliminate the resistance of the boundary between

the solid wall and liquid ^{3}He. (As was shown in the experiments of Lee and Fairbank (1957), which will be discussed below, this resistance exists in both ^{4}He and ^{3}He.) In the apparatus of Challis and Wilks a temperature difference due to the presence of a heat flow was measured at the edges of the smooth cylindrical layer of liquid ^{3}He, which was contained between two copper blocks, bounded at the sides by a thin-walled german-silver tube. Challis and Wilks consider that their values are reduced by 10%, due to the cause mentioned above.

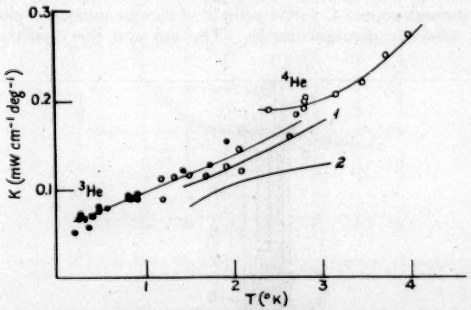


Figure 29. Temperature dependence of thermal conductivity in liquid ^{3}He and ^{4}He (Fairbank and Lee 1957). 1, Thermal conductivity of liquid ^{3}He (Challis and Wilks 1957); 2, thermal conductivity of gaseous ^{3}He (Challis and Wilks 1957).

These data are shown in figure 29 by a solid line 1. Taking into account the corrections for the resistance of boundary layers between the liquid and the solid state, they agree well with the results of Lee, Donnelly and Fairbank. It should be pointed out, however, that although Challis and Wilks consider Fairbank and Lee's values more reliable, these are in fact not sufficiently convincing because firstly, there is no certainty that in Fairbank and Lee's apparatus convection currents were eliminated completely, and secondly, corrections which were introduced for the heat flow in the walls of the tube were numerically greater than the heat flow in the liquid itself.

Besides the thermal conductivity of liquid ^{3}He, Challis and Wilks also measured that of the gas, with the same apparatus. Results of these measurements are also given in figure 29 by a solid line 2. It can easily be seen that the thermal conductivities of liquid and gaseous ^{3}He are very similar at high temperatures, which suggests consideration of the transport phenomena in ^{3}He in the high-temperature region from the point of view of the gas-kinetic theory. Indeed, in the temperature region above 1°K, ^{3}He closely follows the classical formula

$$\kappa = \tfrac{5}{2} C_v \eta \qquad \ldots\ldots(24)$$

where κ is the thermal conductivity, C_v the specific heat, and η the viscosity of liquid ^{3}He. The value of κ calculated from this formula on the basis of known experimental values for C_v (Roberts and Sydoriak 1955) and η (Taylor and Dash

1957) agree sufficiently closely with the directly measured values. Formula (24) ceases to be valid below 1°K.

§ 14. Temperature Discontinuity on the Boundary of ^3He and the Wall

As has already been pointed out by Fairbank and Lee (1957), in experiments on the thermal conductivity of ^3He a temperature discontinuity (jump) appears at the boundary between a solid surface and liquid ^3He, causing dissipation of heat. Investigation of the temperature jump was made with the apparatus for measuring thermal conductivity, with small modifications. Liquid ^3He was condensed in a thin-walled (Cu–Ni) tube 1·5 mm in diameter, which was closed at both ends by pure copper plugs. A thermometer was fixed to the upper copper

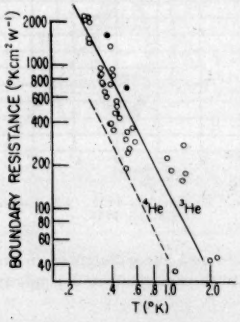


Figure 30. Dependence on temperature of heat resistance at the boundary of ^3He–solid body.

plug and two other thermometers were fixed on the surface of the german-silver (Cu–Ni) tube. A pellet of paramagnetic salt in contact with the lower copper plug acted as a fourth thermometer. The temperature dependence of the resistance of the boundary determined from the data of Fairbank and Lee is shown on a logarithmic scale in figure 30. For comparison, analogous data for ^4He taken from a paper by Fairbank and Wilks (1955) are shown in the same figure. It can easily be seen that the temperature dependence of the resistance boundary of the solid is the same with ^3He and ^4He ; the difference is only in absolute magnitude. The continuous line on the graph corresponds to $R = 130/T^2$ for ^3He and the broken line $R = 45/T^2$ for ^4He. Thus the existence of a temperature discontinuity on the boundary of ^3He and the solid surface, when heat is being transferred across the boundary, shows that this property is inherent in all substances.

§ 15. Surface Tension

The surface tension of ^3He was measured practically simultaneously by Eselson and Bereznyak (1954), Zinov'eva (1955) and Lovejoy (1955), using the same method of capillary rise. Results of these measurements are given in figure 31.

The limiting value of the surface tension $\alpha_0 = 0{\cdot}152$ dyn cm^{-1} is reached at $T = 0{\cdot}6°$K for ^{3}He. This is 2·3 times smaller than the limiting value of α_0 for ^{4}He. Zinov'eva (1955) has shown that the temperature dependences of the surface tension of ^{3}He and ^{4}He are identical. This can easily be seen by comparing the graphs of dependence of α/α_0 on T/T_c for ^{3}He and ^{4}He. The two curves coincide to within the limits of errors of measurement. Thus the difference of statistics

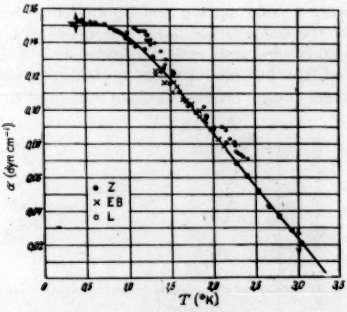


Figure 31. Surface tension of ^{3}He. Z, Zinov'eva 1955; EB, Eselson and Bereznyak 1954; L, Lovejoy 1955.

of the two isotopes of helium, and the differences in states which follow from this, do not lead to a change in the temperature dependence of the surface tension of the liquids.

§16. Velocity of Sound

In 1957 Flicker and Atkins (1958) and simultaneously Laquer, Sydoriak and Roberts (1958) measured the velocity of sound in liquid ^{3}He.

Flicker and Atkins carried out measurements in the temperature interval from 1·05° to 2·2°K with impulses of 14 μsec duration. They used an electronic circuit, analogous to that described by Atkins and Chase (1951). The sound was generated by a quartz crystal, placed on the bottom of a chamber of volume 0·12 cm^3 at a distance 1·013 cm from the reflector. In the temperature region investigated, the speed of sound increased from 145 m sec^{-1} at 2·2°K to 170 m sec^{-1} at 1·0°K. The error of measurement is not greater than $\pm 5\%$.

Laquer, Sydoriak and Roberts (1958) measured the speed of sound in a much wider temperature interval from 0·37° to 3·14°K, using the apparatus shown in figure 32. Measurements were carried out using the impulse method, with a frequency of 5 Mc/s. The sound was generated by the oscillations of the quartz which was situated at the bottom of a thin-walled brass cylinder 1·0447 cm long. The oscillations were produced with a repetition frequency of 7 per sec and had a duration of 0·4 μsec. A second crystal fixed above acted as a sound receiver. The received signal, after amplification by a wide band amplifier, was displayed

on an oscilloscope. By means of a special electronic device, the delay time of the received signal relative to the radiated signal was determined with great accuracy. The temperature of the liquid was determined from the vapour pressure of ^3He. Values of the velocity of sound above 2°K were obtained by the

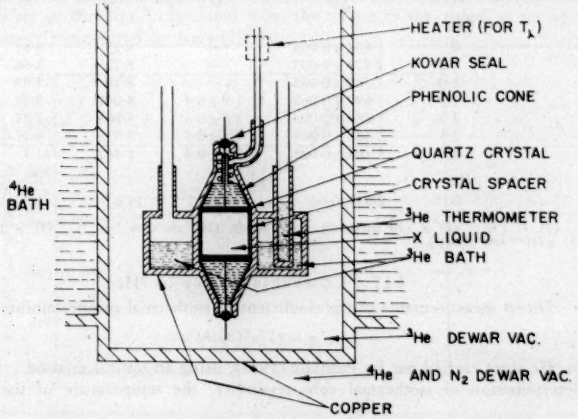


Figure 32. Apparatus for measuring the speed of sound in liquid ^3He (Laquer, Sydoriak and Roberts 1958).

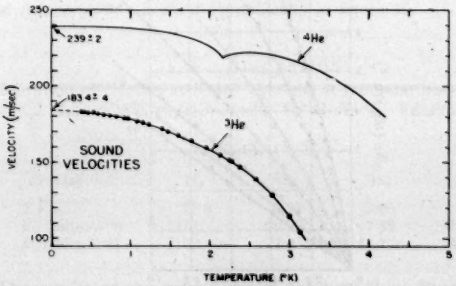


Figure 33. Temperature dependence of velocity of sound in liquid ^3He (Laquer, Sydoriak and Roberts 1958).

extrapolation of results of measurements at higher pressures to saturated vapour pressure. In figure 33 the temperature dependence of the speed of sound, as determined by Laquer, Sydoriak and Roberts, is shown graphically. The mean values of the speed of sound found by these authors (1958) are shown in table 13,

together with an estimate of error of measurement. Also included in this table are values of du/dp and the adiabatic compressibility χ.

Table 13. Sound Velocities and Adiabatic Compressibilities in ^{3}He

(1)	(2)	(3)	(4)	(5)
^{3}He				
0	1·834 ± 0·005	—	$8{\cdot}23_5$	3·61
0·5	1·824 ± 0·003	—	8·22	3·66
1·0	1·780 ± 0·003	—	8·19	3·85
1·5	1·697 ± 0·003	1·9 ± 0·4	8·061	4·31
2·0	1·584 ± 0·003	2·5 ± 0·3	7·805	5·11
2·5	1·421 ± 0·004	4·5 ± 0·5	7·335	6·75
3·0	1·148 ± 0·005	9·8 ± 0·8	6·462	11·7
^{4}He				
0	2·39 ± 0·02	—	14·6	1·20

(1) T (°K); (2) u (10^4 cm sec^{-1}); (3) du/dp (10^{-3} cm^3 dyn^{-1} sec^{-1}); (4) ρ (10^{-3} g cm^{-3}); (5) χ (10^{-3} cm^3 dyn^{-1}).

§ 17. Compressibility of ^{3}He

Direct measurements of the coefficient of isothermal compressibility

$$\chi = (1/\rho)(\partial\rho/\partial p)_T$$

in ^{3}He were carried out by Peshkov (1957), using an optical method. During the determination of isothermal compressibility, the temperature of the bath was

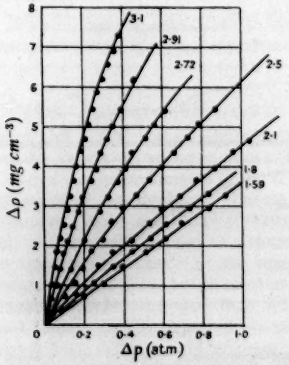


Figure 34. Curves of isothermal compressibility of liquid ^{3}He.
Figures on the curves give temperatures in °K. Experimental points : Peshkov (1957).

maintained constant, and the displacement of the interference bands was observed as the pressure was increased inside the optical chamber. The results of Peshkov's measurements for liquid ^{3}He are plotted in figure 34.

Besides direct measurements of the isothermal compressibility, indirect calculations of the coefficient of adiabatic compressibility of the liquid have been made by Flicker and Atkins (1958), and simultaneously by Laquer, Sydoriak and Roberts (1958) based on measurements of the speed of sound in liquid ^{3}He. Table 13 gives the data of Laquer, Sydoriak and Roberts (1958) on the adiabatic compressibility of the liquid calculated from the values of the speed of sound u and the density ρ measured by Kerr (1954).

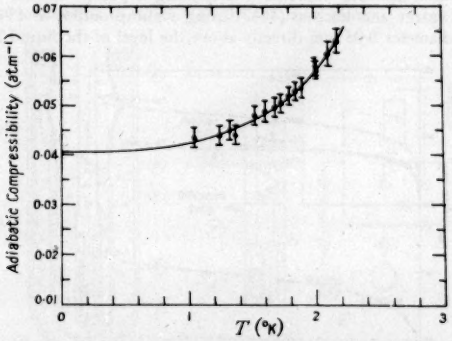


Figure 35. Adiabatic compressibility of liquid ^{3}He (Flicker and Atkins 1958).

The coefficient of adiabatic compressibility of liquid ^{3}He, as calculated by Flicker and Atkins (1958), is shown graphically in figure 35.

§ 18. Adsorption

The adsorption of ^{3}He on activated charcoal at temperatures $T_1 = 2{\cdot}45°$K and $T_2 = 3{\cdot}0°$K was determined experimentally by Hoffman, Edeskuty and Hammel

Table 14

Gas	N_2	^{4}He	^{3}He	^{3}He
T (°K)	75·6	3·95	3·00	2·45
E_L (cal mol^{-1})	1330	20·95	7·88	10·59
E_1 (cal mol^{-1})	2100	71·2	45·3	42·4

(1956). The results of the measurements are plotted in figure 36, where for comparison the data for ^{4}He and N_2 are also given. The measurements show that the volume of adsorbed gas needed for the formation of a monolayer is the same for both isotopes of helium, and is 0·43 cm_0^3 (S.T.P./m^2) for ^{3}He and 0·45 cm_0^3 (S.T.P./m^2) for ^{4}He. Hoffman *et al.* (1956) demonstrated the absence of capillary condensation of ^{3}He at pressures less than boiling pressure. The heat of adsorption E_1 and the latent E_L of the helium isotopes, and nitrogen, are given in table 14.

§19. Cryostats with ^{3}He

The use of ^{3}He instead of ^{4}He to obtain low temperatures by the pumping of liquid helium vapours has a number of advantages. At 1°K the vapour pressure of ^{3}He is 35 times greater than that of ^{4}He, and at 0·5°K it is 10 000 times greater. Further, the existence of a superfluid film in ^{4}He, which creeps up the walls of a Dewar and evaporates at the top, makes it very difficult to obtain extremely low temperatures by pumping. The lowest temperature obtained by pumping ^{4}He gas is 0·71°K, which Keesom obtained (1932) using a pump with an output of 675 l./sec, and Lazarev and Eselson (1942) using a pump output of 15 l./sec and a diaphragm of diameter 0·05 mm directly above the level of the liquid helium.

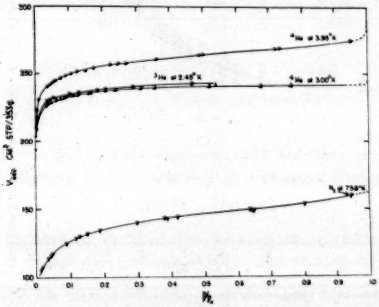


Figure 36. Adsorption of ^{3}He, ^{4}He and N_2 on activated charcoal.

The greatest difficulty in working with ^{3}He is the necessity of using it in small quantities and of ensuring its recovery. However, there are undoubted advantages in using ^{3}He to obtain temperatures below 1°K, and these stimulated the development of cryostats, using ^{3}He as a working substance, in numerous laboratories (Sydoriak and Roberts 1957, Roberts and Sydoriak 1954, Reich and Garwin 1958, Seidel and Keesom 1957, Peshkov, Zinov'eva and Filimonov 1959). The cryostat working at the Institute of Physical Problems in Moscow (Peshkov, Zinov'eva and Filimonov 1959) enables visual observations to be made at temperatures down to 0·3°K. The apparatus is shown in figure 37 to the left, and the diagram on the right. The cryostat consists of two Dewar vessels 1 containing ^{4}He, within which is a small transparent glass Dewar 2 with liquid ^{3}He, of volume about 3 cm^3. ^{3}He gas is pumped through a thin-walled stainless-steel tube 3, 15 mm in diameter, connected to the small Dewar by means of a copper junction 4. To shield the interior of the copper junction from radiation from above, two screens 5 were fixed, in contact with the bath of liquid ^{4}He.

The vapour was pumped with a mercury diffusion pump DRN-50 (figure 37), with an output of 30 l./sec, working at a backing pressure of 25–30 mm Hg. A trap cooled by liquid nitrogen L was fixed on the inlet of the diffusion pump DRN-50. The outlet of the pump was connected to a coil 6 immersed in a bath of liquid ^{4}He, and connected to the small Dewar by a valve 7. The coil passes

successively through the outer and inner Dewars with ^4He, in which temperatures of 1·5° and 1°K respectively are maintained by pumping with separate pumps. Due to the low temperature of the ^4He bath, the mercury pump DRN-50 can be used without a backing pump; at the output it has a pressure approximately equal to ^3He vapour pressure at 1°K, 8–10 mm Hg. ^3He, condensed and cooled in the coil to 1°K, is let into the small Dewar 2 as required by means of the valve 7. In the case of high temperatures of the ^4He bath, and also for the recovery of ^3He into the apparatus, a mercury mechanical pump (TP), developed by Danilov (1956) was connected in series with the diffusion pump.

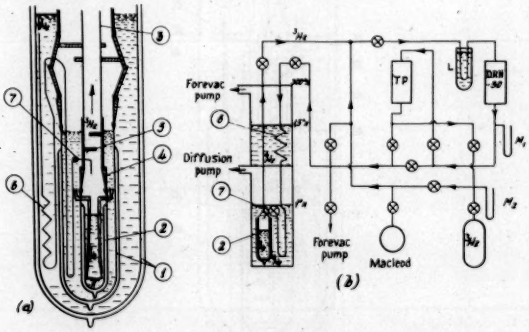


Figure 37. Glass cryostat with ^3He and scheme of apparatus.

A temperature $T = 0{\cdot}3°K$ ($p = 0{\cdot}002$ mm Hg) was reached, using this apparatus. Temperatures of liquid ^3He were determined by a phosphor bronze resistance thermometer, diameter 30 μ, calibrated by the vapour pressure of ^3He and the magnetic susceptibility of chrome potassium alum. Since it was possible to let in liquid ^3He periodically, work on this apparatus at low temperature could continue for a period of 8–10 hours on the same supply of gaseous ^3He (about 3 litres). The temperature 0·5°K was reached 5–10 minutes after letting in. When the pump was switched off the heating of liquid from 0·3° to 1°K took no less than 3 hours.

Another type of apparatus which is in use at the Institute of Physical Problems to obtain low temperatures over large volumes is shown in figure 38. The layout of the apparatus is analogous to that shown in figure 37. The Toepler pump (TP) was replaced by an oil pump NVG-2 with an output of 320 l./hr and backing pressure of 1 atmosphere, developed in NIVI (1955) and distinguished by the fact that it was hermetically sealed. Gaseous ^3He is condensed in the coil 1 (external diameter 2·0 mm, internal diameter 1·4 mm, length 5 m), placed in a bath of liquid ^4He and taken off through a control valve 2 and a coil 3 in the ^3He vessel 4. The coil 3 was made from a capillary of internal diameter 0·2 mm and length 200 mm. The ^3He flowing through the coil 3 is additionally cooled by the vapour evaporated from the liquid below. The casing of the valve 2, being in

contact with the outer bath, also acts as a screen against radiation from above. The ^{3}He vessel 4 is a copper beaker welded to a thin-walled stainless-steel tube, length 100 mm, external diameter 12 mm, internal diameter 11·6 mm. The upper end of the tube is welded to the body of the apparatus. To the bottom of the ^{3}He vessel a copper strip is welded in the form of a spiral 5, with a height 5 mm and thickness 0·3 mm with a 0·2 mm pitch. To obtain optimum heat transport from the walls of the vessel to the coldest layer of ^{3}He, i.e. the surface, it is necessary to increase the surface of the vessel in contact with the boiling ^{3}He.

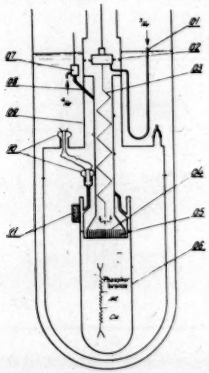


Figure 38. Metal cryostat with ^{3}He, continuous action.
Note. 01,02, . . . are referred to in the text as 1,2 . . .

A copper bulb 6 of capacity about 200 cm^3 was welded to the ^{3}He vessel and liquid ^{4}He from the bath let into it through the valve 7 and the german-silver tube 8 (external and internal diameters 2 mm and 1·4 mm, length 160 mm). The ^{3}He vessel and the bulb are insulated from the helium bath by a vacuum jacket 9, and a mesh 11 impregnated with activated charcoal (1·5 g) is attached to the outside wall of the bulb. There are 12 electrical conductors through the vacuum jacket. The jacket is then filled with ^{4}He under a pressure of 0·5 mm Hg at room temperature and sealed off. The presence of the ^{4}He in the jacket decreases the cooling time of the bulb and the ^{3}He vessel during the starting time of the ^{3}He cryostat. At a known low temperature the ^{4}He in the jacket is absorbed by the charcoal. The time needed to cool the bulb of the cryostat from $T = 4{\cdot}2°\text{K}$ to $T = 0{\cdot}5°\text{K}$ was less than one hour. The time needed to cool the ^{3}He vessel and the bulb with liquid ^{4}He from $T = 1{\cdot}15°\text{K}$ to $T = 0{\cdot}5°\text{K}$ was 5–10 minutes. By regulating the speed of pumping gaseous ^{3}He and its input rate through the valve 2 into the vessel, the temperature of the small bulb could be stabilized at 0·5°K or above. When there is no liquid ^{4}He in the bulb, the temperature can be lowered to 0·35°K. The main heat leak to the bulb is from the bath through the helium inside the tube 8. The temperature inside the bulb 6 is measured

by a phosphor bronze resistance thermometer. The superconducting wires, one of pure aluminium 50 μ in diameter, and the other pure cadmium 67 μ in diameter, are connected in series with a thermometer and serve as calibration points, with respect to which the thermometer is calibrated each time.

Sydoriak and Roberts (1957, Roberts and Sydoriak 1954) used a cryostat for measuring the heat properties and vapour pressure of ^3He, shown in figure 6.

The Dewar with ^3He which was placed in a bath of ^4He, was entirely of metal. ^3He was condensed in a seamless copper sphere of diameter 12·7 mm. Also inside the sphere were crystals of paramagnetic salt for measuring the temperature.

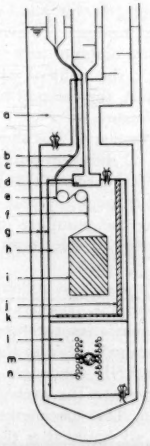


Figure 39. Cryostat for measuring specific heat.

Primary and secondary coils were wound directly on to the brass vacuum jacket. For heat insulation the copper sphere was fixed on to the brass block of the apparatus by an alloy pipe, 70% Cu, 30% Ni, 200 mm long, and 3 mm in diameter, with walls 0·25 mm thick. Different inserts could be screwed into the brass block for measuring the vapour pressure, specific heat, etc., as desired. In the diagram the inserts for measuring vapour pressure are shown. Temperatures below 1°K were obtained by pumping gaseous ^3He. The lowest temperature reached by the authors with this apparatus was 0·37°K.

For measuring the specific heat in the temperature region from 0·3°K, Seidel and Keesom (1958) developed a special type of cryostat, shown in figure 39. It consists of a Dewar a with liquid ^4He inside which is a vacuum jacket g, and a copper vessel, consisting of three different compartments d, h, l. Liquid ^3He condensed in the collector d is pumped through a german-silver tube c, diameter 4 mm and walls 0·1 mm thick. A specimen i, with a carbon thermometer and a

heater on it, is suspended inside the compartment h by a nylon thread, attached to a steel wire f of diameter 0·12 mm. The wire passes up through a system of blocks e and the capillary b of diameter 1 mm. By means of the wire the sample could be suspended in adiabatic conditions, or could be lowered, and thus be in thermal contact with the copper plate k. The plate was connected by a copper rod and a copper jacket to the ^{3}He bath. To accelerate the cooling of the specimen inside the compartment h, traces of air were left behind, which ensured rapid cooling to 77°K. At lower temperatures the air froze and ensured good heat insulation. To measure the temperature inside the compartment l, a powdered paramagnetic salt, ferro-ammonium alum, was sealed off in a glass sphere m, in an atmosphere of ^{4}He (at $p = 1$ atm and $T = 77$°K). Measurement was carried out with an a.c. bridge. To avoid heating by Foucault currents, the primary coil n was made from niobium wire (superconducting below 8°K) and placed inside the compartment l, the walls of which were superconducting 50% Sn, 50 % Pb. For thermal contact, the compartment l was filled with ^{4}He at room temperature and atmospheric pressure.

The ^{3}He pumping system is similar to that shown in figure 37, but the mercury pump has been replaced by an oil diffusion pump of output 25 l./sec with an oil rotary backing pump of output 1·3 l./sec. The gas condensation took place not through the special capillary but through the pumping channel. The temperature was regulated not by measuring the pumping rate, as in the earlier apparatus, but by feeding a regulated quantity of ^{3}He from the collector on to the input of the diffusion pump.

Work with this equipment proceeds as follows. After cooling the apparatus to 77°K, liquid helium is let into the Dewar and is maintained at atmospheric pressure until the specimen reaches a temperature of 4·2°K (from 8 to 24 hours). Later the remains of ^{4}He are pumped out from the vacuum jacket g, the temperature of the bath drops to 1°K and ^{3}He condenses in the container d. The rotary pump is then switched off, and when the pressure in d falls to 0·5 mm Hg the diffusion pump is switched off. The specimen standing on the copper plate k cools to 0·35°K over a period of six hours after the temperature in the collector d reaches 0·3°K. There is no further temperature drop because the thermal contact between the specimen and the copper is very weak at these temperatures, and a heat production in the carbon resistance thermometer of 0·1 erg sec^{-1} produces a temperature difference of 0·05°K. This situation also enables the specific heat of the specimen to be measured without raising it, because upon raising it heat is produced ; this would heat the specimen by tenths of a degree, and fluctuations of temperature caused by the mechanical vibrations would arise. Above 1°K the heat insulation was insufficient and the specimen had to be raised. In figure 40 are shown the temperatures produced in the apparatus, as a function of heat production. With the minimum of heat production of the order of 100 erg sec^{-1} and a supply of $3l_0$ of ^{3}He in the container (at $T = 0$°C and $p = 760$ mm Hg), temperatures around 0·3°K can be maintained continuously for a period of 80 hours.

As is evident from the description, a quantity of 1–3 l_0 of ^{3}He is quite sufficient for the efficient running of the ^{3}He cryostat at temperatures below 1°K. With relatively small heat input, using either the first or the second modifications of the

apparatus described, and a bath of ^4He at 1°K, one can use the mercury diffusion booster pump, and reduce the quantity of ^3He necessary to 0·05–0·1 l_0.

One such cryostat, using 0·05 l_0 in all, was described by Garwin and Reich at the Delft Conference (1958). The advantages of using cryostats with ^3He for work in the temperature region from 0·4° to 1°K, compared with methods using paramagnetic salt or ^4He, are quite clear.

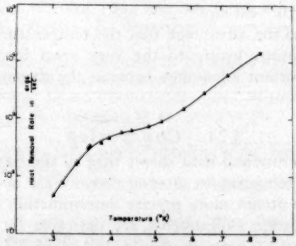


Figure 40. Curve of warming liquid ^3He (T, q).

§ 20. Low Temperature Production by Crystallization of ^3He

As mentioned earlier, Pomeranchuk (1950) suggested that while in liquid ^3He spin alignment begins at temperatures below 1°K, in the solid state such alignment begins even at $T = 10^{-7}$ deg K. Primakoff (1957) considers that in solid ^3He antiparallel alignment begins at $T = 0{\cdot}5$°K ; however, in this case the entropy of the solid helium is probably greater than the entropy of liquid ^3He. On the basis of present experimental data it is impossible to say which of the propositions is correct. The experiments of Walters and Fairbank (1957) showed that the melting of solid ^3He results at $T = 1{\cdot}2$°K in cooling, and at 0·2° in heating ^3He, i.e. in the vicinity of 0·2° the entropy of the solid helium is greater than that of liquid ^3He ; therefore it is in this region that, in principle, it is possible to lower the temperature by adiabatic compression of helium. If Pomeranchuk's suggestions are correct then, by converting liquid ^3He into the solid state by adiabatic compression, it would be theoretically possible to reach temperatures of the order of 10^{-7}–10^{-6} deg K ; however, the amount of cooling by this is very small. In fact, as soon as solid ^3He obtains heat sufficient to increase its entropy to $R\ln 2$, the temperature will rise sharply by tenths of a degree. Because heat is supplied at a level of 10^{-7}–10^{-6} deg K, to heat one gramme atom of ^3He requires only $(R\ln 2) \times 10^{-6} \sim 70$ erg mol^{-1} or about 3 erg cm^{-3}. It is difficult to imagine a compression process which would not generate this amount of heat by internal friction.

The possibility should be borne in mind of constructing a thermostat for very low temperatures embodying the principle of increase of melting pressure of ^3He with decrease of temperature, and consequently adsorption of heat in the process of crystallization. If one begins to increase pressure adiabatically in liquid ^3He cooled to a temperature 0·2°–0·3°K, then when melting pressure is reached,

crystallization will begin and considerable heat adsorption will result. The mixture of solid and liquid ^{3}He begins to cool and the temperature will be stabilized at a value corresponding to equilibrium at a given pressure. Then by adjusting the compression to maintain a constant pressure, it is possible to maintain a constant temperature, the heat input being compensated by the heat of crystallization. In this case the supply of cold will be greater than the higher regulating temperature, and the order of magnitude will be equal to ΔST_p, i.e. at $T_p \sim 10^{-3}$ deg K, about 10^{-3} cal mol^{-1}.

This process still has the advantage that the temperature will be constant and uniform within wide limits, owing to the very great intermingling of the two phases; this is an important advantage because the thermal conductivity of ^{3}He is very small.

§21. Conclusion

A review of the experimental data shows that all the basic properties of liquid and gaseous ^{3}He in the temperature interval above 0·2°K are known, and the only remaining problem is to obtain more precise determination of various parameters. In particular, more accurate experiments are desirable for the determination of the thermal conductivity of gaseous, liquid and solid ^{3}He. The properties of solid ^{3}He are not fully known. It would be interesting to obtain data on the variation of the transition curve from the α to the β phase in ^{3}He for all temperature ranges of the existing phases, and also data on the measurement of volume and heat content at transition and the magnetic properties of the phases.

The properties of liquid and solid ^{3}He at temperatures below 0·2°K are of great interest. The temperature dependence of entropy in solid ^{3}He in this region is unknown and by its behaviour may be determined the character of the spin interaction of ^{3}He nuclei, which is of great importance in examining the possibility of using the crystallization of ^{3}He by pressure, for obtaining and maintaining very low temperatures.

Below 0·05°K the properties of ^{3}He can be described by Landau's (1956) theory concerning the spectral excitation of Fermi fluids. This problem is considered further in the survey of Abrikosov and Khalatnikov (1959).

A point of peculiar interest is in question about the superfluidity of ^{3}He. Like electrons in superconductors, ^{3}He atoms, because of their nuclear spins, have a tendency towards marked pair interaction, therefore at very low temperatures one could expect to find superfluidity in ^{3}He. The solution of this problem is of fundamental importance.

Acknowledgments

The authors express their deep gratitude to Professor Strelkov for reading the manuscript and giving much valuable advice, and acknowledge permission to reproduce various figures and tables.

References

Abraham, B. M., Osborne, D. W., and Weinstock, B., 1949, *Phys. Rev.*, **75**, 988; 1950, *Ibid.*, **80**, 366; 1955, *Ibid.*, **98**, 551.

Abrikosov, A. A., and Khalatnikov, I. M., 1956, *J. Exp. Theor. Phys., U.S.S.R.*, **32**, 1083. (see *Soviet Physics, J.E.T.P.*, **5**, 887); 1958, *Uspecki Phys. Nauk*, **66**, 177.

Aldrich, I. L. T., and Nier, A. O., 1948, *Phys. Rev.*, **74**, 1590.

Anderson, H. L., and Novik, A., 1948, *Phys. Rev.*, **73**, 919.

Atkins, K. R., and Chase, C. E., 1951, *Proc. Phys. Soc.* A, **64**, 826.

Becker, E. W., and Misenta, R., 1954, *Phys. Rev.*, **93**, 244.

Beenakker, J., Taconis, K., Lynton, E., Dokoupil, Z., and Vam Soest, G., 1952, *Physica*, **18**, 433.

Bloembergen, N., Purcell, E. M., and Pound, R. V., 1948, *Phys. Rev.*, **73**, 679.

de Boer, J., and Cohen, E. G. D., 1951, *Physica*, **17**, 993.

de Boer, J., Van Krakendonk, J., and Compaan, K., 1950, *Physica*, **16**, 545.

Brewer, D. F., Sreedhar, A. K., Kramers, H. C., and Daunt, J. G., 1958, *Phys. Rev.*, **110**, 282.

Brueckner, K. A., and Gammel, J. L., 1957, *Proceedings of the Symposium on Solid and Liquid ^3He*, Ohio State University, p. 186.

Challis, L. J., and Wilks, J., 1957, *Proceedings of the Symposium on Solid and Liquid ^3He*, Ohio State University, p. 38.

Clement, J. R., Logan, J. K., and Gaffney, J., 1955, *Phys, Rev.*, **100**, 743.

Danilov, I. B., 1956, *P.T.E. USSR*, No. 1, p. 93.

Dash, J., and Taylor, R., 1955, *Phys. Rev.*, **99**, 598.

Daunt, J. G., Probst, R. E., and Johnston, H. L., 1947 a, *J. Chem. Phys.*, **15**, 759; 1948, *Phys. Rev.*, **73**, 638.

Daunt, J. G., Probst, R. E., Johnston, H. L., Aldrich, L. T., and Nier, A. O., 1947 b, *Phys. Rev.*, **72**, 502.

Daunt, J. G., Probst, R. E., and Smith, S., 1948, *Phys. Rev.*, **74**, 495.

Eselson, B. N., and Bereznyak, N. G., 1954, *Dokl. Akad. Nauk S.S.S.R.*, **99**, 365; 1956, *J. Exp. Theor. Phys., U.S.S.R.*, **30**, 628.

Eselson, B. N., Bereznyak, N. G., and Kaganov, M. I., 1956, *Dokl. Akad. Nauk S.S.S.R.*, **111**, 568.

Eselson, B. N., and Lazarev, B. G., 1950, *J. Exp. Theor. Phys., U.S.S.R.*, **20**, 742; 1954, *Dokl. Akad. Nauk, USSR.*, **97**, 61.

Eselson, B. N., Lazarev, B. G., and Alexeevski, N. E., 1950, *J. Exp. Theor. Phys., U.S.S.R.*, **20**, 1055.

Fairbank, H. A., and Lee, D. M., 1957, *Proceedings of the Symposium on Solid and Liquid ^3He*, Ohio State University, p. 27, 32.

Fairbank, H. A., and Wilks, J., 1955, *Proc. Roy. Soc.* A, 231, 545.

Fairbank, W. M., and Adams, E. D., 1958, *Kamerlingh Onnes Conference*, Leiden.

Fairbank, W. M., Ard, W. B., Dehmelt, H. G., Gordy, W., and Williams, S. R., 1953, *Phys. Rev.*, **92**, 208.

Fairbank, W. M., Ard., W. B., and Walters, G. K., 1954, *Phys. Rev.*, **95**, 566.

Fairbank, W. M., and Walters, G. K., 1957 a, *Proceedings of the Symposium on Solid and Liquid ^3He*, Ohio State University, p. 205; 1957 b, *Ibid.*, p. 221.

Flicker, H., and Atkins, R. K., 1958, *Proceedings of the Symposium on Solid and Liquid ^3He*, Ohio State University, p. 11.

Frank, I., 1946, *Phys. Rev.*, **70**, 561.

Garwin, R. L., Kan, —., and Reich, H. A., 1955, *N.S.F. Conference on Low Temperatures*, Barton Rouge L.

Garwin, R. L., and Reich, H. A., 1958, *Kamerlingh Onnes Conference, Leiden*, p. 25.

Grilly, E. R., and Mills, R. L., 1958, *Proc. 5th Internat. Conf. on Low Temperature Physics and Chemistry, Wisconsin.*

Hammel, E. F., Laquer, H. L., Sydoriak, S. G., and McGee, W. E., 1952, *Phys. Rev.*, **86**, 432.

Hammel, E. F., Sherman, R. H., Kilpatrick, J. E., and Edeskuty, F. J., 1958, *Kamerlingh Onnes Conference, Leiden.*

HEER, C. V., and DAUNT, J. G., 1949, *Phys. Rev.*, **76**, 854.
HOFFMAN, C. J., EDESKUTY, E. J., and HAMMEL, E. F., 1956, *J. Chem. Phys.*, **24**, 124.
HOLMES, J. E. H., 1949, *Proc. Phys. Soc.* A, **62**, 293.
KEESOM, W. H., 1932, *Proc. Roy. Acad. Amsterdam*, **35**, 136.
KELLER, W. E., 1955, *Phys. Rev.*, **98**, 1571.
KERR, E. C., 1954, *Phys. Rev.*, **96**, 551.
KHALATNIKOV, I. M., and ABRIKOSOV, A. A., 1957, *J. Exp. Theor. Phys., U.S.S.R.*, **32**, 915 (see *Soviet Physics, J.E.T.P.*, **5**, 745).
KILPATRICK, J. E., KELLER, W. E., and HAMMEL, E. F., 1955, *Phys. Rev.*, **97**, 9.
KILPATRICK, J. E., KELLER, W. E., HAMMEL, E. F., and METROPOLIS, N., 1954, *Phys. Rev.*, **94**, 1103.
KISTEMAKER, J., 1946, *Physica*, **12**, 272, 281.
KUZNETSOV, V. M., 1957, *J. Exp. Theor. Phys., U.S.S.R.*, **32**, 1001 (see *Soviet Physics, J.E.T.P.*, **5**, 819).
LANDAU, L. D., 1956, *J. Exp. Theor. Phys., U.S.S.R.*, **30**, 1058.
LANE, C. T., FAIRBANK, H. A., ALDRICH, L. T., and NIER, A. O., 1948, *Phys. Rev.*, **73**, 256.
LAQUER, H. L., SYDORIAK, S. G., and ROBERTS, T. R., 1957, *Proceedings of the Symposium on Liquid and Solid ^3He*, Ohio State University, p.15.
LAZAREV, B. G., and ESELSON, B. N., 1942, *J. Exp. Theor. Phys., U.S.S.R.*, **12**, 549.
LEE, D. M., DONNELLY, R. J., and FAIRBANK, H. A., 1957, *Bull. Amer. Phys. Soc. II*, **2**, 64.
LONDON, F., and RICE, O. K., 1948, *Phys. Rev.*, **73**, 1188.
LOVEJOY, D. R., 1955, *Canad. J. Phys.*, **33**, 49.
MCINTYRE, B. B., ALDRICH, L. T., and NIER, A. O., 1948, *Phys. Rev.*, **74**, 946.
MILLS, R. L., and GRILLY, E. R., 1955, *Phys. Rev.*, **99**, 480.
NIVI, 1955, *Symposium on the exchange of advanced work in the fields of technology and organization, MRTP, S.S.S.R.*, p. 55.
OSBORNE, D. W., ABRAHAM, B. M., and WEINSTOCK, B., 1951, *Phys. Rev.*, **82**, 263; 1954, *Ibid.*, **94**, 202.
OSBORNE, D. W., and WEINSTOCK, B., 1958, *Kamerlingh Onnes Conference on Low Temperature Physics, Leiden.*
PESHKOV, V. P., 1956 a, *J. Tech. Phys. U.S.S.R.*, **26**, 665; 1956 b, *J. Exp. Theor. Phys., U.S.S.R.*, **30**, 850; 1957, *Ibid.*, **33**, 833 (see *Soviet Physics, J.E.T.P.*, **6**, 645).
PESHKOV, V. P., and KACHINSKI, V. N., 1956, *J. Exp. Theor. Phys., U.S.S.R.*, **31**, 720 (see *Soviet Physics, J.E.T.P.*, **4**, 607).
PESHKOV, V. P., ZINOV'EVA, K. N., and FILIMONOV, A. I., 1959, *J. Exp. Theor. Phys., U.S.S.R.*, in the press.
POLLARD, E., and DAVIDSON, W., 1945, *Applied Nuclear Physics.*
POMERANCHUK, I. YA., 1949, *J. Exp. Theor. Phys., U.S.S.R.*, **19**, 42; 1950, *Ibid.*, **20**, 919.
PRIMAKOFF, H., 1957, *Bull. Amer. Phys. Soc. Ser. II*, **2**, 63.
PTUKHA, T. P., 1958, *J. Exp. Theor. Phys., U.S.S.R.*, **34**, 33 (see *Soviet Physics, J.E.T.P.*, **7**, 22).
REICH, H. A., and GARWIN, R. L., 1958, *Proceedings of 1st Conference of Institut International du Froid, Delft.*
ROBERTS, T. R., and SYDORIAK, S. G., 1954, *Phys. Rev.*, **93**, 1418; 1955, *Ibid.*, **98**, 1672; 1956, *Ibid.*, **102**, 304; 1958, *Proceedings of 5th International Conference on Low Temperature Physics and Chemistry, Wisconsin.*
SEIDEL, G., and KEESOM, P. H., 1958, *Proceedings of the Fifth International Conference on Low Temperature Physics and Chemistry, Wisconsin*, p. 218.
SHERMAN, R. H., and EDESKUTY, F. J., 1957, *Proceedings of the Conference on Solid and Liquid ^3He*, Ohio State University, p. 44.
SOLLER, T., FAIRBANK, W. M., and CROWELL, A. D., 1953, *Phys. Rev.*, **91**, 1058.
SOMMERS, H., 1952, *Phys. Rev.*, **88**, 113.
SYDORIAK, S. G., GRILLY, E. R., and HAMMEL, E. F., 1949, *Phys. Rev.*, **75**, 303, 1103.
SYDORIAK, S. G., and HAMMEL, E. F., 1949, *Proceedings of the International Conference on Physics of Very Low Temperatures*, M.I.T., Cambridge, p. 42.
SYDORIAK, S. G., and ROBERTS, T. R., 1957, *Phys. Rev.*, **106**, 175.

TACONIS, K. W., 1950, *Ned. Tijdschr. Natuurk.*, **16**, 101.
TAYLOR, R. D., and DASH, J. G., 1957, *Phys. Rev.*, **106**, 398.
TAYLOR, R. D., and KERR, E. C., 1958, *Kamerlingh Onnes Conference on Low Temperature Physics*, Leiden, p. 25.
TISZA, L., 1948, *Physics Today*, **1**, 26.
VAN DIJK, H., and DURIEUX, M., 1956, *Physics*, **22**, 760; *Conference on Physics of Low Temperatures*, p. 599.
VAN DIJK, H., and SHOENBERG, D., 1949, *Nature, Lond.*, **164**, 151.
DE VRIES, G., and DAUNT, J. G., 1953, *Phys. Rev.*, **92**, 1572; 1954, *Ibid.*, **93**, 631.
WALTERS, G. K., and FAIRBANK, W. M., 1956 a, *Phys. Rev.*, **103**, 262(L); 1956 b, *Ibid.*, **103**, 263(L); 1957, *Bull. Amer. Phys. Soc., Ser. II*, **2**, 183.
WEINSTOCK, B., ABRAHAM, B. M., and OSBORNE, D. W., 1952, *Phys. Rev.*, **85**, 158.
WEINSTOCK, B., OSBORNE, D. W., and ABRAHAM, B. M., 1950, *Phys. Rev.*, **77**, 400.
ZAVOISKY, E., 1945, *J. Phys.*, **9**, 211.
ZINOV'EVA, K. N., 1955, *J. Exp. Theor. Phys., U.S.S.R.*, **29**, 899; 1958, *Ibid.*, **34**, 609 (see *Soviet Physics, J.E.T.P.*, **7**, 421).
ZINOV'EVA, K. N., and PESHKOV, V. P., 1959, *J. Exp. Theor. Phys., U.S.S.R.*, **36**, no. 4.

THE TWO-NUCLEON INTERACTION

By R. J. N. PHILLIPS
Atomic Energy Research Establishment, Harwell, Berks

CONTENTS

	PAGE
§ 1. General introduction	563
§ 2. Phenomenology	564
2.1. Introduction	564
2.2. The form of the potential	565
2.3. The description of scattering	567
2.4. Analysis in terms of potentials	571
2.4.1. Evidence from the deuteron	571
2.4.2. Evidence from other nuclei	572
2.4.3. Low-energy scattering	574
2.4.4. High-energy scattering	577
2.4.5. Other experiments	581
2.4.6. Spin–orbit coupling	583
2.4.7. Separable potentials	584
2.5. The boundary condition model	584
2.6. Phase-shift analysis	586
2.7. The construction of potentials from phase shifts	591
2.8. The Gammel–Thaler potential	594
2.9. Outlook	597
§ 3. Meson field theory	597
3.1. Introduction	597
3.2. The construction of a potential	599
3.2.1. The canonical transformation method	599
3.2.2. The $\mathscr{S}$-matrix method	602
3.2.3. The Tamm–Dancoff method	603
3.2.4. The Lévy–Klein method	606
3.2.5. The Brueckner–Watson method	608
3.2.6. Uniqueness	609
3.3. Potentials from the pion field	610
3.3.1. The pion field	610
3.3.2. The form of the potential	612
3.3.3. The phenomenological inner region	613
3.3.4. The one-pion exchange potential	613
3.3.5. The two-pion exchange potential	614
3.3.6. Static potential models	616
3.3.7. Correction terms	620
3.3.8. Pair suppression	623
3.3.9. Spin–orbit coupling	624
3.3.10. The Signell–Marshak potential	625
3.3.11. Dispersion relations	626
3.4. Outlook	627
Acknowledgments	627
References	627

Abstract. In recent years there have been great advances in understanding the interaction of two nucleons, both phenomenologically and in terms of a fundamental meson–nucleon coupling. The present report describes the methods and achievements of phenomenology and meson field theory in this subject.

§1. General Introduction

This paper is a survey of the progress made in recent years toward finding and explaining the forces between two non-relativistic nucleons. Separate chapters deal with the two main approaches to this subject: phenomenology and meson field theory.

The phenomenological approach aims to interpret the observed properties of the interaction in terms of some model—usually a potential—and thus to gain an empirical understanding of them. One hopes that a more fundamental understanding will follow.

The meson field theory of nuclear forces attempts a fundamental explanation, based on Yukawa's original suggestion that the interaction may occur through an intermediary field associated with particles of about three hundred electron masses. The pion (or π-meson), the only light meson with strong nuclear interactions, is now thought to be responsible for much of the interaction, although the heavier mesons about which less is known may contribute significantly at very short distances.

The success of phenomenological potentials in explaining low-energy experiments was extended to high-energy scattering cross sections by Jastrow's 'hard core' potential model in 1950. After many vicissitudes, meson field theory began to offer an explanation of at least the low-energy data, with the work of Taketani, Lévy and others from 1951 onwards. However, neither approach found an explanation for the various polarization measurements that became possible in high-energy scattering, until 1957. Early in that year two qualitatively similar potentials were proposed, each of which gave a fair explanation of all the data up to quite high energies. One of these was constructed phenomenologically by Gammel and Thaler; the other, due to Signell and Marshak, was derived in part from pion field theory. These simultaneous successes for both approaches mark the peak of present achievement.

These successes are not complete. The agreement with data is fair in most respects, but far from perfect. In the construction of the Gammel–Thaler potential there is much arbitrariness which only more experimental information can remove. Many points in the Signell–Marshak potential also are arbitrary, uncertain or arguable. Nevertheless it appears that a real advance in understanding the two-nucleon interaction has been made; this seems an appropriate moment to review the situation.

The chapter on phenomenology is meant to be intelligible to any reader with a working knowledge of quantum mechanics, such as may be acquired from Schiff (1949), Mandl (1957) or other standard textbooks. A grounding in elementary nuclear theory (e.g. Bethe and Morrison 1956) would be an advantage.

The chapter on meson field theory presupposes a familiarity with modern (post-1949) field theory; Schweber, Bethe and de Hoffmann (1955) is a suitable textbook. Unfortunately this subject cannot be treated in simpler terms without sacrificing much of its content, but the general reader should be able to gather many of the ideas and results from §§ 3.1, 3.2.6, 3.3.2, 3.3.3, 3.3.10, 3.3.11, 3.4, and perhaps some other sections.

The bibliography is extensive. This and the condensed summaries of results that occur at intervals in the text are intended for the serious student; others may safely ignore them.

The present review was completed, apart from minor corrections, at the end of August, 1958.

The reader may like to refer to earlier reviews of the two-nucleon problem by Hulthén and Sugawara (1957) and by a number of authors in the *Progress of Theoretical Physics* Supplement No. 3 (1956). The latter deal purely with pion field theory and emphasize the work of the Japanese school, sometimes overlooked in the West. A book on nuclear scattering by Mather and Swan (1958), with many relevant chapters, has recently appeared. A review of two-nucleon phenomenology by Gammel and Thaler (1959), contemporary with the present article, will also appear soon.

§ 2. Phenomenology

2.1. *Introduction*

This chapter describes the phenomenological approach to the two-nucleon problem through the non-relativistic Schrödinger equation. A large number of experiments have been and will be performed to discover the properties of the two-nucleon interaction; it is important to have at least an empirical understanding of these properties in terms of some interaction model. The models investigated are usually potentials, but may include (or consist purely of) boundary conditions on the wave function. We shall review the successes and failures of various models.

The opening sections are mainly introductory. Section 2.2 explains the language and notation of potential models and the particular forms usually chosen. Section 2.3 introduces the $\mathscr{S}$-matrix, phase shifts and the description of scattering experiments.

The following sections describe the information derived from nuclear structure and scattering experiments, and its interpretation in terms of a potential or a boundary condition model. Scattering data, particularly that at high energies, are so complex that attempts to explain them all directly in terms of some model have little chance of success; however, great insight can be gained by first analysing the data in terms of phase shifts. Section 2.6 is concerned with such analyses and the results they give.

If phase shifts are sufficiently well known, as functions of energy, it may be possible to deduce a potential directly; § 2.7 deals with this question. However, even a limited knowledge can be very useful; a few known phase shifts at a single high energy were the starting-point of Gammel and Thaler in constructing their successful potential model (§ 2.8).

The final section, § 2.9, contains some remarks on the future outlook.

Throughout this chapter, explicit mathematics is avoided and well-known results are mentioned as briefly as possible. Details can be found in the papers cited. For long-established results, the reader is often referred for convenience to the textbooks by Blatt and Weisskopf (1952) and Sachs (1953), or the Encyclopaedia article by Hulthén and Sugawara (1957), rather than the original papers.

As to notation, we shall use units in which $\hbar = c = 1$. In scattering problems the energy quoted is that of the bombarding particle in the laboratory system; scattering angles are given in the centre-of-mass system.

2.2. *The Form of the Potential*

We consider two interacting non-relativistic nucleons with coordinates $\mathbf{r}^{(i)}$, momenta $\mathbf{p}^{(i)}$ and Pauli spin operators $\boldsymbol{\sigma}^{(i)}$ $(i = 1, 2)$. We treat the nucleon charge-states, neutron and proton, like the spin-states of a spin-$\frac{1}{2}$ particle and introduce the usual isobaric (or 'isotopic') spin operators $\boldsymbol{\tau}^{(i)}$. Proton and neutron are eigenstates of τ_3, with eigenvalues $+1$ and -1 respectively; the matrix elements of τ_1 and τ_2 couple these states. The total charge of the system is $1 + T_3$, where T is the total isobaric spin.

The Schrödinger equation in the centre-of-mass frame is

$$(\mathbf{p}^2/M + V)\psi = i d\psi/dt \qquad \ldots\ldots(1)$$

where $\mathbf{p} = \frac{1}{2}(\mathbf{p}^{(1)} - \mathbf{p}^{(2)})$ is the relative momentum, M is the nucleon mass, and V is the interaction potential, a function of the $\mathbf{p}^{(i)}$, $\mathbf{r}^{(i)}$, $\boldsymbol{\sigma}^{(i)}$ and $\boldsymbol{\tau}^{(i)}$. Galilean and translational invariance require V to depend on the $\mathbf{p}^{(i)}$ and $\mathbf{r}^{(i)}$ only through $\mathbf{p}$ and $\mathbf{r} = \mathbf{r}^{(1)} - \mathbf{r}^{(2)}$. Equation (1) appears like the Schrödinger equation of a single particle of mass $\frac{1}{2}M$, interacting through V with a fixed centre of force.

If V depends on $\mathbf{p}$ as well as $\mathbf{r}$ it is 'non-static' or 'non-local'. In the $\mathbf{r}$-representation it is non-diagonal, and may be represented as an integral operator (e.g. $(V\psi)(\mathbf{r}') = \int (\mathbf{r}'|V|\mathbf{r}'')\psi(\mathbf{r}'')\,d\mathbf{r}''$). On the other hand, we may write $\mathbf{p} = -i\boldsymbol{\nabla}$ and treat V as a differential operator.

We wish to deduce V from observed properties of interacting nucleons. However, we remark at once that nucleon–nucleon scattering and the deuteron binding energy cannot fix it uniquely. Any canonical transformation of the Hamiltonian $\mathbf{p}^2/M + V$ which leaves the asymptotic form of the wave function unchanged provides a new Hamiltonian $\mathbf{p}^2/M + V'$ with the same eigenvalues and scattering properties. We may try to fix V by considering electromagnetic properties of the deuteron, for different potentials give different wave functions, but this too involves some arbitrariness. Definite assumptions must be made about the contributions of virtual mesons; we need a complete field theory to treat this question consistently. Similarly, we cannot invoke many-body phenomena without making definite assumptions about possible many-body potentials. In practice, V is usually fixed by making some assumptions about its form, for example† that it is static.

The potential is generally taken to have the following properties:

(i) Invariance under rotations in space.

(ii) Invariance under reflections in space (parity conservation). Once taken for granted, recently questioned, this is now confirmed experimentally to a high degree (Tanner 1957, Wilkinson 1958, Jones *et al.* 1958).

(iii) Invariance under time-reversal. This is also confirmed by experiment, though less precisely (Hillman *et al.* 1958, Abashian and Hafner 1958, Rosen and Brolley 1959, Bodansky *et al.* 1959).

† This assumption is too severe to fit arbitrary data, but is useful in certain contexts.

(iv) Invariance under rotations in isobaric space (charge independence). This means that V is the same for any pair of nucleons, apart from electromagnetic effects and the exclusion principle (which forbids certain states for identical particles). It is supported by experiment to a fair approximation (§§ 2.4.2, 2.4.3, 2.4.4).

A further assumption is often made for convenience:

(v) V is at most linear in $\mathbf{p}$.

Assuming (i)–(v), Eisenbud and Wigner (1941) show that V can only depend on $\mathbf{r}$, $\mathbf{p}$, $\boldsymbol{\sigma}^{(i)}$ and $\boldsymbol{\tau}^{(i)}$ in the following ways:

$$V = V(r), \qquad \ldots\ldots(2a)$$

$$V = \boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}\, V(r), \qquad \ldots\ldots(2b)$$

$$V = S_{12}\, V(r) \equiv \{3(\boldsymbol{\sigma}^{(1)}.\mathbf{r})(\boldsymbol{\sigma}^{(2)}.\mathbf{r})/r^2 - \boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}\}\, V(r), \qquad \ldots\ldots(2c)$$

$$V = \mathbf{L}.\mathbf{S} V(r) \equiv \tfrac{1}{2}\mathbf{r}\times\mathbf{p}.(\boldsymbol{\sigma}^{(1)}+\boldsymbol{\sigma}^{(2)})\, V(r), \qquad \ldots\ldots(2d)$$

or any of the above multiplied by $\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}$, or any linear combination of these. (2*a*) is a static central potential; (2*b*) is spin-dependent, $\boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}$ having eigenvalues $+1$ and -3 in triplet and singlet states, respectively; (2*c*) is a non-central 'tensor' potential; (2*d*) is a non-central, non-static 'spin–orbit' potential.

Okubo and Marshak (1958) give the general form of V when all powers of $\mathbf{p}$ are allowed.

It is conventional, and loses no generality, to use wave functions that are antisymmetric under the interchange of all nucleon coordinates in space, spin and isobaric spin. Now the operator $P_\sigma = \frac{1}{2}(1+\boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)})$ has eigenvalues $+1$ and -1 for symmetric (triplet) and antisymmetric (singlet) spin states, respectively, and may thus be regarded as interchanging the nucleon spins. Similarly, $P_\tau = 1+\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}$ exchanges isobaric spins (or charges). From the total antisymmetry of the wave function, it follows that $P_x = -P_\sigma P_\tau$ exchanges space coordinates. The dependence of V on $\boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}$ and $\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}$ is often discussed in terms of these exchange operators. Central potentials of pure spin-exchange, charge-exchange, space-exchange and non-exchange character are known as Bartlett, Heisenberg, Majorana and Wigner potentials, respectively.

The radial dependence ('shape') of V is usually discussed in terms of trial functions; the most common are:

$$V(r) = \begin{cases} V_0, & r < r_0 \\ 0, & r > r_0 \end{cases} \qquad \text{('square well');} \qquad \ldots\ldots(3a)$$

$$V(r) = V_0 \exp(-r/r_0)^2 \qquad \text{('Gaussian well');} \qquad \ldots\ldots(3b)$$

$$V(r) = V_0[1-\exp(-r/r_0)]^{-1}\exp(-r/r_0) \qquad \text{('Hulthén well');} \qquad \ldots\ldots(3c)$$

$$V(r) = V_0(r_0/r)\exp(-r/r_0) \qquad \text{('Yukawa well');} \qquad \ldots\ldots(3d)$$

$$V(r) = V_0 \exp(-r/r_0) \qquad \text{('exponential well').} \qquad \ldots\ldots(3e)$$

V_0 and r_0 are depth and range parameters. The Yukawa well is suggested by meson field theory, which predicts this shape for the central potential in lowest order (however, the important non-central and higher-order terms have different

shapes, see § 3.3). The Hulthén well is similar to the Yukawa well, and gives a soluble Schrödinger equation for S-states and central forces. In recent years, potential models have usually contained a 'hard core' (infinitely repulsive region) at small distances, suggested by high-energy scattering results (§ 2.4.4).

Instead of assumption (v) above, the potential is sometimes taken to be 'separable', that is $(\mathbf{r}'|V|\mathbf{r}'') = \lambda w^*(\mathbf{r}')v(\mathbf{r}'')$, $(\mathbf{p}'|V|\mathbf{p}'') = \lambda g^*(\mathbf{p}')g(\mathbf{p}'')$, suppressing the dependence on spin and isobaric spin for simplicity. This type of non-local interaction gives a soluble Schrödinger equation and has other practical advantages (Yamaguchi 1954). If $g(\mathbf{p}) = g(p)$, the potential is central and acts only in S-states. If we allow a dependence on the direction of the vector $\mathbf{p}$, and take a sum of separable terms, the potential given by

$$(\mathbf{p}'|V|\mathbf{p}'') = \lambda\Sigma_m Y_L^{m*}(\theta', \phi')Y_L^m(\theta'', \phi'')g^*(p')g(p'')$$

(in an obvious notation) can be shown to act only in the state of angular momentum L. This too gives a soluble Schrödinger equation. Non-central separable potentials can be constructed by allowing terms such as $\boldsymbol{\sigma}.\mathbf{p}$ in $g(\mathbf{p})$.

2.3. *The Description of Scattering*

The $\mathscr{S}$-matrix. We have essentially to describe scattering by a fixed centre of force, according to (1). At a given energy, a complete set of quantum numbers are the total spin S, the total charge $1+T_3$, the orbital angular momentum about the scattering centre L, the total angular momentum J and the component J_Z; all are conserved when $V=0$. Let $\phi(LSJJ_Z\,T_3)$ be a complete orthonormal set of solutions of the Schrödinger equation with $V=0$, labelled by these quantum numbers; each can be separated into ingoing and outgoing parts, ϕ_{in} and ϕ_{out}.

Assuming that $V\to 0$ rapidly enough as $r\to\infty$, we can separate the asymptotic part of any solution ψ of (1) into ingoing and outgoing parts, and make an $LSJJ_Z\,T_3$ representation of each part in terms of the sets ϕ_{in} and ϕ_{out}. The asymptotic ingoing and outgoing parts of ψ are related by the familiar $\mathscr{S}$-matrix†

$$\psi(\text{asymp. out}) = \mathscr{S}\psi(\text{asymp. in}). \qquad \ldots\ldots(4)$$

We thus set up an $LSJJ_Z\,T_3$ representation of $\mathscr{S}$; this representation is convenient for introducing basic parameters (see *phase shifts*, below).

Conservation of probability makes $\mathscr{S}$ unitary; $\mathscr{S}^*\mathscr{S}=1$. Invariance under rotations in space and isobaric space makes the matrix elements independent of J_Z and T_3. With suitable conventions for the relative phases of the set ϕ (in fact, the usual ones: Condon and Shortley 1935), time-reversal invariance makes $\mathscr{S}$ symmetric. Blatt and Weisskopf (1952) and Sachs (1953) give simple discussions of $\mathscr{S}$, with references.

The scattered wave, as distinct from the total outgoing wave, is determined by the matrix $\mathscr{S}-1$; it is to this that experimental cross sections are most closely related. To describe scattering experiments, it is convenient to introduce a $\mathbf{p}\sigma_3^{(i)}\tau_3^{(i)}$ representation (replacing the ϕ above by free orthonormal eigenfunctions of $\mathbf{p}, \tau_3^{(i)}$ and $\sigma_3^{(i)}$), and then to write the matrix elements of $\mathscr{S}-1$ with respect to $\mathbf{p}$ as functions of initial and final relative momenta $\mathbf{k}_i$ and $\mathbf{k}_f$, and the operators $\boldsymbol{\sigma}^{(i)}$ and $\boldsymbol{\tau}^{(i)}$. These matrix elements are usually denoted by M. Assuming

† Strictly, this is only the submatrix of $\mathscr{S}$ for a given energy.

charge-independence and invariance under rotations, reflections and time-reversal,† the most general form is

$$M = A + B(\sigma_N^{(1)} + \sigma_N^{(2)}) + C\sigma_N^{(1)}\sigma_N^{(2)} + D\sigma_P^{(1)}\sigma_P^{(2)} + E\sigma_K^{(1)}\sigma_K^{(2)} \quad \ldots\ldots(5)$$

(Dalitz 1952, Wolfenstein and Ashkin 1952). **N**, **P** and **K** are unit vectors in the directions $\mathbf{k}_i \times \mathbf{k}_f$, $\mathbf{k}_i + \mathbf{k}_f$ and $\mathbf{k}_f - \mathbf{k}_i$, respectively, and $\sigma_N^{(1)} = \boldsymbol{\sigma}^{(1)}.\mathbf{N}$, etc. The coefficients $A, B, \ldots$, etc. are complex functions of energy, scattering angle and T (equivalently, of $k_i^2 = k_f^2 = k^2$, $\mathbf{k}_i.\mathbf{k}_f$ and $\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}$). They are not completely independent; the unitarity of $\mathscr{S}$ imposes five conditions, relating the coefficients at any angle to integrals over all angles. Hence M and $\mathscr{S}$ may be determined in principle at a given energy by five independent experiments at all angles for each value of T (Puzikov *et al.* 1957).

We define the polarization of a nucleon beam by the statistical expectation value of the spin operator, $\langle\boldsymbol{\sigma}\rangle$. The state of polarization of such a beam may be described by a 2×2 density matrix ρ, for which the most general form is $\rho = \frac{1}{2}(1 + \mathfrak{P}.\boldsymbol{\sigma})$. Since $\langle\boldsymbol{\sigma}\rangle \equiv \text{trace } \rho\boldsymbol{\sigma}/\text{trace } \rho = \mathfrak{P}$, we identify $\mathfrak{P}$ as the beam polarization. For an unpolarized beam, $\mathfrak{P} = 0$; for a fully polarized beam $|\mathfrak{P}| = 1$.

To describe the state of polarization of a two-nucleon system, we need a 4×4 density matrix, for which the most general form is

$$\rho = \tfrac{1}{4}(1 + \mathfrak{P}^{(1)}.\boldsymbol{\sigma}^{(1)} + \mathfrak{P}^{(2)}.\boldsymbol{\sigma}^{(2)} + \textstyle\sum_{ij} C_{ij}\sigma_i^{(1)}\sigma_j^{(2)}).$$

Since $\langle\boldsymbol{\sigma}^{(i)}\rangle \equiv \text{trace } (\rho\boldsymbol{\sigma}^{(i)})/\text{trace } \rho = \mathfrak{P}^{(i)}$, we identify the $\mathfrak{P}^{(i)}$ as the individual nucleon polarizations. The coefficients $C_{ij} = \langle\sigma_i^{(1)}\sigma_j^{(2)}\rangle$ describe possible correlations between the spins.

In a scattering experiment, if the initial density matrix is ρ_i that of the scattered wave is $\rho_f = M\rho_i M^*$. We now see how experimental measurements are related to M. If the initial system is unpolarized and uncorrelated ($\rho_i = \frac{1}{4}$) the differential cross section is trace ρ_f/trace $\rho_i = \frac{1}{4}$ trace MM^*, and the final polarization of nucleon i is $\langle\boldsymbol{\sigma}^{(i)}\rangle = \text{trace } MM^*\boldsymbol{\sigma}^{(i)}/\text{trace } MM^*$. If initially nucleon i has polarization $\mathfrak{P}^{(i)}$, the cross section is $\frac{1}{4}$ trace $(MM^* + M\mathfrak{P}^{(i)}.\boldsymbol{\sigma}^{(i)}M^*)$, and so on.

This non-relativistic polarization formalism is given by Dalitz (1952) and Wolfenstein and Ashkin (1952). At relativistic energies nucleons must be treated as Dirac particles; the appropriate theory of polarization phenomena is developed by Stapp (1955, 1956).

Phase shifts. If V is such that the quantum numbers L and S are conserved, $\mathscr{S}$ is diagonal in the $LSJJ_Z T_3$ representation. Since $\mathscr{S}$ is unitary, the diagonal elements are simply phase factors $\exp 2i\delta_{LSJ}$, independent of J_Z and T_3 for reasons already given. The parameters δ_{LSJ} are called phase shifts; a more common notation is δ_L for $S = 0$ and $\delta_L{}^J$ for $S = 1$. A zero phase shift means that the corresponding partial wave is not scattered.

L is not conserved in general, however (e.g. when V contains a tensor terms $(2c)$). $\mathscr{S}$ is always diagonal with respect to the spin singlet states and the triplet states with $L = J$ by virtue of its invariance properties, but may couple the triplet

† For the generalization of (5) and phase shifts without time-reversal invariance, see Mandl (1958) and Phillips (1958).

states with $L = J \pm 1$. With respect to each such pair of states, $\mathscr{S}$ is a unitary and symmetric 2×2 matrix, which can therefore be written

$$\mathscr{S}(L = J \pm 1) = U^{-1} \Delta U, \qquad \ldots\ldots(6)$$

$$U = \begin{pmatrix} \cos \epsilon^J, & \sin \epsilon^J \\ -\sin \epsilon^J, & \cos \epsilon^J \end{pmatrix}, \qquad \ldots\ldots(7)$$

$$\Delta = \begin{pmatrix} \exp 2i\delta_\alpha{}^J, & 0 \\ 0, & \exp 2i\delta_\beta{}^J \end{pmatrix}, \qquad \ldots\ldots(8)$$

(Blatt and Biedenharn 1952 a, b). It is thus characterized by three real parameters. In effect the transformation U diagonalizes $\mathscr{S}$; Δ represents $\mathscr{S}$ in terms of its eigenstates. $\delta_\alpha{}^J$ and $\delta_\beta{}^J$ are called 'eigenphase shifts'; they are often conventionally written $\delta_L{}^J$ in terms of the L-value taken when $\epsilon^J \to 0$.† ϵ^J is called the 'mixture angle'.

Stapp (1955, Stapp *et al.* 1957) suggests the alternative parametrization,

$$\mathscr{S}(L = J \pm 1) = \begin{pmatrix} \exp i\bar{\delta}_{J+1}{}^J, & 0, \\ 0, & \exp i\bar{\delta}_{J-1}{}^J \end{pmatrix} \begin{pmatrix} \cos 2\bar{\epsilon}^J, & i \sin 2\bar{\epsilon}^J \\ i \sin 2\bar{\epsilon}^J, & \cos 2\bar{\epsilon}^J \end{pmatrix} \begin{pmatrix} \exp i\bar{\delta}_{J+1}{}^J, & 0 \\ 0, & \exp i\bar{\delta}_{J-1}{}^J \end{pmatrix}. \qquad \ldots\ldots(9)$$

The 'bar' parameters $\bar{\delta}_{J\pm1}{}^J$ and $\bar{\epsilon}^J$ are not so simple to interpret as the others, but have advantages when Coulomb effects are discussed. They are also more directly related to the potential, in Born approximation.

We shall henceforth refer to the basic $\mathscr{S}$-matrix parameters (including mixture angles and bar parameters) collectively as 'phase shifts'. It is a straightforward matter to express cross sections and other observed quantities in terms of phase shifts; this is done by many authors, including Ashkin and Wu (1948), Breit and Hull (1955), Breit *et al.* (1955), Stapp (1955), Wright (1955) and Stapp *et al.* (1957).

Because nuclear forces have a short range, we expect only a limited number of partial waves (with L sufficiently small) to be significant in scattering at a given energy. Classically, the trajectories for large L do not come within range of the potential and are not scattered; there is an analogous quantum-mechanical argument.

Modification for a Coulomb potential. We have hitherto assumed that $V \to 0$ rapidly enough as $r \to \infty$ for the nucleons to behave asymptotically like free particles. This is true for specifically nuclear forces, which are characteristically short-ranged, but not for the unscreened Coulomb‡ potential e^2/r which appears in the p–p interaction.

Whereas the free wave functions $\phi(LSJJ_Z)$ depend asymptotically on r as $\mp r^{-1} \exp \pm i(kr - \frac{1}{2}\pi L)$ (the upper sign for outgoing, the lower sign for ingoing

† The sign of ϵ^J depends on the correspondence here. Blatt and Biedenharn let α correspond to $L = J-1$. Some authors use the opposite convention.

‡ There is always some screening by atomic electrons in practice, but it is awkward to include explicitly and can be neglected at the energies and scattering angles usually considered.

parts), the corresponding Coulomb wave functions $\phi^C(LSJJ_Z)$ behave asymptotically as $\mp r^{-1}\exp\pm i(kr-\frac{1}{2}\pi L+\sigma_L-\eta\ln 2kr)$, where k is the relative momentum, $\eta=e^2M/2k$ and the σ_L are 'Coulomb phase shifts'.

The usual procedure is to modify the sets ϕ_{in} and ϕ_{out} by the progressive phase factors $\pm\eta\ln 2kr$, and define $LSJJ_Z$ representations of the ingoing and outgoing parts of ψ in terms of these new sets. A modified $\mathscr{S}$-matrix is thus defined, denoted $\mathscr{S}'$, with the same properties as before (strictly, we are concerned only with the $T_3=1$ submatrix). For pure Coulomb scattering, $\mathscr{S}'$ is diagonal with elements $\exp 2i\sigma_L$.

If a nuclear potential which conserves L and S is introduced, $\mathscr{S}'$ remains diagonal, and the elements are written $\exp 2i(\sigma_L+\delta_L')$ and $\exp 2i(\sigma_L+\delta_L{}^{J'})$ for singlet and triplet states, respectively. The δ_L' and $\delta_L{}^J$ are not the ordinary phase shifts that would be found if the Coulomb field were absent; they are called 'quasi-nuclear' phase shifts, and we distinguish them by a prime (the latter is often omitted in the literature). However, at high energies the differences between the two can usually be neglected for practical purposes, with plausible potential models.†

In general $\mathscr{S}'$ is not diagonal in the $LSJJ_Z$ representation, but couples the states $L=J\pm1$. We write the 2×2 submatrix of $\mathscr{S}'$ in the form (9) replacing $\bar{\epsilon}^J$ by $\bar{\epsilon}^{J'}$ and $\delta_{J\pm1}{}^J$ by $\delta_{J\pm1}{}^{J'}+\sigma_{J\pm1}$, thus defining the quasi-nuclear bar parameters.

This choice of parameters is readily understood. The most natural way to describe specifically nuclear deviations from pure Coulomb scattering is to define a matrix $\mathscr{S}_N$ in analogy with $\mathscr{S}$, but using a complete orthonormal set of Coulomb wave functions $\phi^C(LSJJ_Z)$ in place of the free wave functions ϕ. $\mathscr{S}_N$ so defined is unitary, symmetric, invariant under rotations and reduces to 1 for pure Coulomb scattering. The basic parameters of $\mathscr{S}_N$ are precisely our quasi-nuclear parameters. The advantage of Stapp's parametrization (9) is that it allows a simple relation between the parameters of $\mathscr{S}'$ and those of $\mathscr{S}_N$.

$\mathscr{S}'$ may also be put in a $\mathbf{p}\sigma_3{}^{(i)}$ representation. However, $\mathbf{p}$ here is not the momentum of a plane wave but labels the appropriate distorted wave.

For more detailed discussions of Coulomb effects, and references, see Mott and Massey (1949) and Blatt and Weisskopf (1952).

Relativistic and other corrections. The definition of phase shifts is easily made relativistic for spinless particles; the free wave functions in the time-independent representation remain the same as in the non-relativistic case, except that $\mathbf{k}$ is now the *relativistic* relative momentum. Relativistic nucleons must be treated as Dirac particles, however. Stapp (1955, 1956, see also Stech 1956 and Chou and Shirokov 1958) examines this question and finds that the non-relativistic spin formalism can be retained, with suitable interpretations, provided that an additional rotation of the polarization vector is added at each scattering. $\sigma(\theta)$ and $P(\theta)$ are not affected, but the definitions of some triple-scattering and other quantities are modified (§ 2.4.4).

Relativistic corrections to Coulomb effects are considered by Garren (1954, 1955, 1956), treated more carefully by Breit (1955, 1957) and Ebel and Hull (1955),

† e.g., with the Gammel–Thaler potential (§ 2.8) above 40 MeV (private communication from J. L. Gammel).

and discussed again by Durand (1957). If the Coulomb–nuclear cross-effect from wave function distortion† can be ignored, the main correction is to re-define η as e^2/v, where v is the relativistic velocity of the incident proton in the laboratory frame.

Garren (1954, 1955, 1956) includes magnetic moment corrections to p–p scattering, in first Born approximation, treating the protons as Dirac particles with anomalous Pauli moments. The effect is not negligible at small angles in polarization experiments, but it is very sensitive to wave function distortion, which is ignored (Breit 1957). Ohnuma (1957) illustrates these relativistic and magnetic moment corrections in sample calculations for high-energy p–p scattering.

Vacuum polarization (Foldy and Eriksen 1954, 1955, Eriksen *et al.* 1956, Durand 1957, de Wit and Durand 1958) gives a correction to the Coulomb phase shifts. This is important at low energies (see § 2.4.3) but can be ignored at high energies for practical purposes.

Breit (1958 b) draws attention to possible effects of atomic electrons, which may not be negligible in accurate low-energy p–p experiments.

2.4. *Analysis in terms of Potentials*

2.4.1. *Evidence from the deuteron.*

The deuteron is the bound n–p state. Experiment shows there are no other bound n–p or p–p states, and if charge independence holds there is no bound n–n state either. The known properties of the deuteron are its binding energy, magnetic dipole moment and electric quadrupole moment (for references to experiments, see Hulthén and Sugawara 1957). The deuteron wave function proves to be mainly $^3S_1$ with a small $^3D_1$ admixture.

The deuteron is usually discussed in terms of static potentials, attributing the S–D mixing to a tensor term. A spin–orbit term is sometimes included; Vaidya (1958) considers the effect on central and tensor parts; the effect on the magnetic moment is discussed below. On the other hand, Yamaguchi and Yamaguchi (1954) use a non-local separable potential (§ 2.4.7).

The binding energy is $2{\cdot}226 \pm 0{\cdot}002$ MeV.

The magnetic dipole moment, $\mu_d = 0{\cdot}857411 \pm 0{\cdot}000019\ e/2M$, where M is specified as the proton mass. It is not exactly the sum of the free neutron and proton moments. If we assume that each nucleon contributes the free moment appropriate to its orientation, and include the proton's orbital moment in the $^3D_1$ state, the discrepancy $\mu_n + \mu_p - \mu_d = 0{\cdot}02231 \pm 0{\cdot}00012$ determines the D-state probability, $P_D \simeq 4\%$.

Now associated with $\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}$-dependent potentials there are certain currents, called 'exchange currents', in the space around interacting nucleons; they are not wholly determined‡ by the potential alone (see Sachs 1953, Foldy 1953, Dalitz 1954). Another type of current is associated with non-static potentials (Sachs 1953), such as a spin–orbit potential. Both types we call 'interaction currents'.

In analysing μ_d we have assumed that the 'interaction moments' contributed by these currents vanish. A more general assumption (often made for exchange moments at least) is that the interaction moments obey a 'mirror condition', being equal and opposite for mirror nuclei (which differ by the interchange

† In empirical phase-shift analyses, this effect is necessarily unknown.
‡ To fix them the underlying field theory must be known and solved.

$n \longleftrightarrow p$). This assumption appears consistent with the deuteron, ^3H and ^3He, and with estimates† for the mirror nuclei of mass 6, 10 and 14.

Simple gauge-invariance arguments for the spin–orbit potential $V(r)\mathbf{L}.\mathbf{S}$ give a non-zero interaction moment‡ for the deuteron (see Sachs 1953, Feshbach 1957). However, Soper (to be published) observes that this potential may be written in other forms, for which gauge invariance arguments give different interaction moments; such arguments are therefore inadequate; the interaction moments may in fact obey the mirror condition.

The information obtained about nuclear forces, namely P_D, is sensitive to small corrections to the analysis; a 1% correction to $\mu_n+\mu_p$ makes a 40% correction to the discrepancy $\mu_n+\mu_p-\mu_d$, and to P_D. Meson field theory indicates that the neutron and proton do not contribute their free magnetic moments when close together (this would be described phenomenologically by interaction moments), and also predicts small relativistic corrections. Certain estimates of these effects, some of which cancel, reduce P_D to about 3% (see Sugawara 1955 a, b, c for details and references) but present meson-theoretical calculations are not wholly reliable. Machida (1953) argues from a comparison of the hydrogen and deuterium hyperfine structure that P should$_D$ be larger. Recent explanations of deuteron photodisintegration at moderate energies (§2.4.5) also require values of the order of 7%. We hesitate to draw a firm conclusion.

The electric quadrupole moment, $Q = 2{\cdot}738 \pm 0{\cdot}014 \times 10^{-27}\, e\,\mathrm{cm}^2$. This gives more information about the wave function. Writing the latter as $[u(r)+w(r)\,S_{12}]/r$ multiplied by a triplet spin function, we find

$$Q = (8\pi e/5)\int_0^\infty (uw - w^2)\, r^2\, dr,$$

with a suitable normalization for u and w. Q is not simply related to

$$P_D = 32\pi \int_0^\infty w^2\, dr.$$

We have ignored possible effects of interaction currents. However, gauge invariance defines the electric multipole operators uniquely, independently of exchange or other effects, provided nucleons are point particles ('Siegert's theorem', see Sachs 1953 and Foldy 1953 for discussion and references). We expect some violation of Siegert's theorem, for nucleon structure is not wholly negligible compared with the size of the deuteron, but the correction here may well be small, for Q is not affected when the proton form factor is spherically symmetrical (Bernstein 1957). Fortunately, small corrections have only a small effect on the predicted property of the wave function (in contrast to P_D).

For more detailed treatments of μ_d and Q, see Blatt and Weisskopf (1952) and Sachs (1953).

2.4.2. *Evidence from other nuclei.*

Charge independence. The energy levels of light nuclei support this hypothesis. Sets of states with the same T but different T_3 can be identified; the energy

† Private communication from J. M. Soper.

‡ This interaction moment has a different effect on the hyperfine structure from a static magnetic moment, making the hyperfine structure an additional independent datum (Sessler and Foley 1958).

levels agree quite well, allowing approximately for Coulomb energy (Lauritsen 1952, Burcham 1955, Wilkinson 1956). This is not confined to mirror nuclei, which would only confirm the weaker 'charge-symmetry' hypothesis (invariance under $n \leftrightarrow p$).

Spin–orbit force in the shell model. The nuclear shell model† in its simplest form supposes nucleons outside a closed shell (and 'holes' in it) to move independently in a central potential plus an $\mathbf{L}.\boldsymbol{\sigma}$ spin–orbit potential. Proton–nucleus scattering experiments also suggest such a spin–orbit coupling (e.g. Sternheimer 1955, Glassgold 1959).

The spin–orbit coupling, which is much larger than the relativistic Thomas correction to the central potential, may come from a spin–orbit term

$$[U(r)+U'(r)\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}]\,\mathbf{L}.\mathbf{S}$$

in the real two-nucleon potential. The nature of this term is not yet firmly established. Elliott and Lane (1954) require $U<0$ and U' relatively small‡; Abraham (1956) estimates $U<0, U'>|U|>0$; Hüper (1957) finds $U\simeq 3U'<0$; Skyrme (1959) suggests $U<0, |U|\gtrsim U'\gtrsim 0$ (these are all essentially first-order estimates). It appears difficult to fit all the data at once; much depends on which are chosen.

Other suggested explanations are a reduced effective mass, amplifying the Thomas term (Johnson and Teller 1955), and three-body forces (e.g. Fujita and Miyazawa 1957). Higher-order effects of two-body tensor forces have been considered, but appear to be too weak (Jancovici 1957, 1958, Brueckner 1958, Nigam and Sundaresan 1958 a).

Chisholm and Touschek (1953) and Laing (1955) offer yet another explanation. The emission and reabsorption of virtual pseudoscalar mesons by a nucleon in a central attractive potential gives a spin–orbit coupling of the right sign as a lowest-order correction. This argues a two-nucleon spin–orbit correction too, of course.

To summarize, a two-nucleon $\mathbf{L}.\mathbf{S}$ term is suggested but not demanded.

Two-body forces in the shell model. As a refinement, the shell model introduces a two-body pseudopotential§ between the nucleons (and holes) considered. It is usually taken to be central, with sometimes a tensor part; the shape proves unimportant. The exchange character is often fixed arbitrarily, but there have been efforts to deduce it empirically (Soper 1957).

Now the one-body central and spin–orbit shell model potentials are in principle determined by a given two-body interaction (neglecting many-body forces). Such a self-consistent approach is made by Hüper (1957) and Skyrme (1959) directly from a two-body pseudo-potential (Skyrme includes a three-body term), for which a general form is assumed.

We cannot expect this pseudo-potential to resemble the real two-nucleon potential closely. However, it can be deduced from the latter in principle, if many-body forces are ignored, and approximate ways of doing so have been proposed (Brueckner 1955, Skyrme 1959). This gives a new though somewhat remote test for two-nucleon potentials. It is interesting that the pseudo-potential

† See Elliott and Lane (1957) for a full description.
‡ Also $U'>0$ preferably: private communication from J. P. Elliott.
§ Treated in first order, it is supposed to represent the effect of real potentials in all orders.

found by Skyrme (1959) is closely related to the Gammel–Thaler and Signell–Marshak potentials (§§ 2.8, 3.3.10), at least in the mean properties (averaged over spin and isobaric spin).

Saturation. The observed saturation of nuclear binding energies was traditionally discussed in terms of simple attractive two-body potentials and gave definite limitations on their exchange character (see Rosenfeld 1948, Blatt and Weisskopf 1952). However, the two-nucleon potential is now believed to contain a repulsive core, and new methods of analysis are needed. Such methods are being developed by Brueckner (1955) and others (for recent papers see Brueckner and Gammel (1958); for an introductory review, see Karplus and Watson (1957)); they may ultimately offer a test of two-body potentials, to the extent that many-body forces can be ignored. An explanation in terms of many-body forces (e.g. Drell and Huang 1953) remains open, and largely unexplored.

2.4.3. *Low-energy scattering.*

For present purposes, low energies are those below 10 MeV.

n–p scattering. We expect only S-wave scattering to be important at sufficiently low energies; possible $^3S_1$–$^3D_1$ coupling does not affect this, for $\epsilon^1 \propto k^2$ in the limit of relative momentum $k \to 0$ (at least for potentials of finite range (Blatt and Biedenharn 1952 a)).

The n–p differential cross section for S-waves is

$$\sigma_{np}(\theta) = (1/4k^2)(\sin^2\delta_0 + 3\sin^2\delta_0^1). \qquad \ldots\ldots(10)$$

It proves convenient to make the expansion, in powers of k^2,

$$\left.\begin{aligned} k\cot\delta_0 &= -1/a_s + \tfrac{1}{2}r_{0s}k^2 + \ldots \\ k\cot\delta_0^1 &= -1/a_t + \tfrac{1}{2}r_{0t}k^2 + \ldots \end{aligned}\right\}. \qquad \ldots\ldots(11)$$

The parameters a and r_0 are called 'scattering lengths' and 'effective ranges', respectively; the suffixes s and t distinguish singlet and triplet. The virtue of these expansions is that the first two terms closely approximate the whole series for the potential models (3) usually studied, at energies small compared with the well depth V_0, and that this approximation describes observed n–p scattering quite well up to about 10 MeV.

Low-energy n–p scattering thus depends on only two parameters for each spin state. Since a potential of any shape contains two adjustable parameters (the depth and range scales), we can fit the data with many different shapes, provided only that they make the higher terms in the expansions (11) small. The third term is usually written $-Pr_0^3k^4$; P is called the 'shape-dependent parameter'. The first two terms alone constitute the 'shape-independent approximation'.

In this approximation, and if the potential has a finite range or tends to 0 sufficiently fast as $r \to \infty$, the deuteron binding energy is related to a_t and r_{0t}. In this case the bound state is associated with the value $\cot\delta = i$; the energy is $-K^2/M$, where $K = 1/a_t + \frac{1}{2}r_{0t}K^2$.

The theory of this approximation and references to the original literature are given in standard works (Blatt and Weisskopf 1952, Sachs 1953). Biedenharn

and Blatt (1954) treat $^3S_1$–$^3D_1$ coupling rigorously, deriving an expansion (11) for the eigenphase shift $\delta_0{}^1$ and analogous expansion for ϵ^1 and $\delta_2{}^1$.

The best shape-independent fits to data are

$$\left.\begin{aligned} a_s &= (-23{\cdot}69 \pm 0{\cdot}055) \times 10^{-13}\ \text{cm} \\ a_t &= (5{\cdot}377 \pm 0{\cdot}021) \times 10^{-13}\ \text{cm} \\ r_{0s} &= (2{\cdot}40 \pm 0{\cdot}23) \times 10^{-13}\ \text{cm} \\ r_{0t} &= (1{\cdot}704 \pm 0{\cdot}028) \times 10^{-13}\ \text{cm} \end{aligned}\right\}. \qquad \ldots\ldots(12)$$

We quote from Hulthén and Sugawara (1957), who describe the experiments and give references.

These, with Q and P_D, are the known low-energy parameters of the n–p system. Just as the scattering is fitted by many shapes of static potential, Q and P_D can also be fitted by adding a tensor potential of ill-determined shape (Feshbach and Schwinger (1951) fit the data with Yukawa shape, and refer to other work). A tensor force affects $^3S_1$-scattering in second order, acting like an attractive central force.

Higher-energy scattering contains information about potential shape. However, partial waves with $L>0$ cannot be ignored here, and it is difficult to extract δ_0 and $\delta_0{}^1$ from the data (see § 2.6). It may prove easier at intermediate energies (<100 MeV, say) to work directly with potential models. Biedenharn *et al.* (1958) begin a systematic investigation by fitting many static central-plus-tensor potentials to the low-energy n–p data. These will be compared with data at successively higher energies, until the static assumption fails.

Raphael (1957) proposes a 'generalized effective range theory' to describe the limitations placed on the potential by a knowledge of phase shifts at higher energies.

p–p scattering. There are two differences from the n–p case. Firstly, we cannot neglect scattering in partial waves with $L>0$, because of the long-ranged Coulomb potential; however, we expect the quasi-nuclear phase shifts to be important in S-waves only. Secondly, antisymmetry excludes $T=0$ states.

The p–p differential cross section is therefore

$$\sigma_{pp}(\theta) = \sigma_C(\theta) + \sigma_{int}(\delta_0'; \theta) + (1/k^2)\sin^2\delta_0', \qquad \ldots\ldots(13)$$

where σ_C is the known Coulomb cross section and σ_{int} is a somewhat complicated term describing Coulomb–nuclear interference.

An effective-range theory can again be set up, with some modification. The quantity most closely analogous to $k\cot\delta$ proves to be

$$K' \equiv C^2 k\cot\delta_0' + Me^2 h(\eta). \qquad \ldots\ldots(14)$$

C^2 is the Coulomb penetration factor $2\pi\eta/(\exp 2\pi\eta - 1)$; $C^2 \to 1$ as $k\to\infty$.

$$h(\eta) = -\ln\eta - \gamma + \eta^2 \sum_1^\infty 1/n(n^2+\eta^2),$$

and $\gamma = 0{\cdot}5772\ldots$ is Euler's constant. In analogy with (11) we can expand

$$K' = -1/a' + \tfrac{1}{2}r_0' k^2 - P' r_0'^3 k^4 + \ldots. \qquad \ldots\ldots(15)$$

a', r_0' and P' are called the p–p scattering length, effective range and shape-dependent parameter (the primes are often omitted). For any given potential shape, the p–p and n–p scattering lengths are related, but there is no shape-independent relation between them. r_0', P' and higher coefficients are much less sensitive to shape, and remain close to their n–p counterparts for reasonable shapes.

For a full treatment of p–p effective range theory, and references, see standard textbooks (Blatt and Weisskopf 1952, Sachs 1953).

The best shape-independent fit to data is

$$\left.\begin{aligned} a' &= (-7{\cdot}694 \pm 0{\cdot}012) \times 10^{-13} \text{ cm} \\ r_0' &= (2{\cdot}656 \pm 0{\cdot}017) \times 10^{-13} \text{ cm} \end{aligned}\right\}. \qquad \ldots\ldots(16)$$

We quote from Hulthén and Sugawara (1957), who give references to the experiments.

Scattering at higher energies contains information about potential shape.† Early cross section data up to 30 Mev display no recognizable P- or D-wave effects; analysed in terms of S-waves, they favour long-tailed potentials of exponential or Yukawa shape (Yovits *et al.* 1952). More recent data reveal higher-wave effects in this region (Johnston and Young 1958, MacGregor 1958), but until the phase shifts are accurately determined a fresh shape-analysis cannot be made.

We can test charge independence by comparing p–p and n–p singlet parameters. The equivalent effective range, which would be found were there no Coulomb potential, differs little from r_0'; r_{0s} and r_0' are consistent with charge independence. The equivalent scattering length is much larger than a' but still considerably less than a_s for all the conventional‡ well shapes (3). However, the scattering length in the present case is very sensitive to changes in well depth, and it seems possible to relate the discrepancy to electromagnetic effects.

Schwinger (1950) shows that the magnetic moment interaction explains this discrepancy, for a Hulthén or Yukawa shape. Salpeter (1953) estimates the effect of a hard core in the potential by introducing a finite repulsive core and supposing the nuclear magnetic moment to be spread over a region the size of this core; for plausible core radii, Schwinger's explanation is destroyed (however, see footnote‡). Vacuum polarization corrections make the agreement—or lack of it—somewhat worse (Foldy and Eriksen 1954, 1955); so does the introduction of nucleon charge and magnetic moment form factors, indicated by electron scattering (Riazuddin 1958).

There is a possible explanation in pion field theory. The mass differences between pions (§ 3.3.1) give small differences between predicted n–p and p–p potentials. Taking some leading terms, Sugie (1954) shows that this difference has the required sign and order of magnitude. Using different pion masses presumably amounts to including some electromagnetic corrections: there would be many others in a complete field theory. Riazuddin (1958) estimates that the lowest order electromagnetic correction to the coupling constant together with the mass correction may fit the discrepancy in scattering lengths.

† Though the accurate data below 5 Mev allow no firm conclusion about shape they appear to suggest $P' > 0$ (Gammel and Thaler 1959). Current hard-core potentials (§§ 2.8, 3.3.10) have $P' < 0$.

‡ The discrepancy vanishes for some less conventional shapes. Preston and Shapiro (1954; Shapiro and Preston 1956) show this for the potential $V = V_0 x^{-3} \exp(-x)$, with a hard core of suitable radius.

The low-energy data thus appear not inconsistent with charge independence, at least for the underlying field theory.

Breit and Bouricius (1949) give an alternative treatment of low-energy n–p and p–p scattering; see § 2.5.

Triplet P-wave anomaly. The most sensitive measure of small ^{3}P phase shifts is a Coulomb–nuclear interference term in $\sigma_{pp}(\theta)$, proportional to

$$\Delta' = \tfrac{1}{9}(\delta_1{}^{0'} + 3\delta_1{}^{1'} + 5\delta_1{}^{2'}).$$

In Born approximation and neglecting P–F coupling, Δ' comes from the central part of the potential; tensor and spin–orbit terms make no contribution. Very accurate measurements of $\sigma_{pp}(\theta)$ from 1·8 to 4·2 Mev show a small but significant departure from S-wave scattering, attributable to a small negative Δ' increasing from about $-0{\cdot}05°$ to $-0{\cdot}1°$ in this interval (Hall and Powell 1953). At 18 Mev, however, Δ' is near $+1{\cdot}0°$ (Yntema and White 1954).

This anomaly can be explained by a potential consisting of a weak repulsive long-ranged part plus a strong attractive short-ranged part (Keller 1955). Pion field theory indicates such a central potential in triplet odd states, consistent with these values of Δ' (Otsuki and Tamagaki 1954, 1955; § 3.3.6).

Part of the effect observed at 1·8–4·2 Mev can be ascribed, however, to vacuum polarization, which effectively contributes $-0{\cdot}047°$ at 1·8 Mev and $-0{\cdot}040°$ at 4·2 Mev to Δ' (Eriksen *et al.* 1956, de Wit and Durand 1958). The nuclear ^{3}P effect is reduced but remains significant; there is no conflict with pion field theory, because of the uncertainties in the latter (§ 3.3.6).

The effects of atomic electrons may not be negligible in this energy region (Breit 1958 b), but quantitative estimates are lacking.

2.4.4. *High-energy scattering.*

We are interested in energies up to about 300 Mev. At higher energies pion production (threshold 290 Mev) cannot be ignored and relativistic effects should become increasingly important; a two-nucleon Schrödinger equation cannot be adequate to describe the interaction.

Types of experiments. Nucleon–nucleon experiments below 40 Mev are at present restricted to unpolarized cross sections, but at higher energies polarization effects are measured.

When unpolarized nucleons scatter through an angle θ, they develop spin polarizations $P(\theta)$ normal to the scattering plane† (parity conservation forbids a component in this plane) through the action of non-central forces. When a nucleon beam of polarization P_i strikes unpolarized target nucleons, a left–right asymmetry $\epsilon = P_i \mathscr{A}(\theta)$ is found in the scattering. Time-reversal and parity arguments show that $\mathscr{A}(\theta) = P(\theta)\cos\phi$, where ϕ is the angle between the incident polarization and the normal to the scattering plane (Dalitz 1952, Wolfenstein and Ashkin 1952); polarization is linked to asymmetry. Similar results hold for elastic nucleon scattering from unpolarized nuclei.

Double-scattering experiments measure the polarization function $P(\theta)$. A nucleon beam acquires polarization $P_1(\theta_1)$ in a first scattering, and develops an asymmetry $P_1(\theta_1)P_2(\theta_2)$ in a second coplanar scattering. If both targets are

† In the notation of § 2.3, $\mathfrak{B} = P(\theta)\mathbf{N}$.

of the same material, $|P_1(\theta)| = |P_2(\theta)|$ may be deduced (ignoring energy degradation); the sign is found separately by comparison with a calculable scattering. Usually P_1 is known in advance, and P_2 deduced.

Triple-scattering experiments measure the polarizations after a nucleon–nucleon collision when one particle is initially polarized. A first scattering produces a beam of known (transverse) polarization which strikes a second, unpolarized target. The final scattering serves to analyse the transverse polarization of the scattered beam or of the recoiling target nucleons. By using magnetic fields to rotate beam polarizations, longitudinal as well as transverse components can be produced and analysed.

Correlation experiments analyse the polarization of both nucleons in coincidence (they measure the coefficients C_{ij} in the density matrix of the scattered system, see § 2.3). More complicated experiments are conceivable, but have not yet been performed.

The theory of scattering and polarization experiments is discussed briefly in § 2.3, and at greater length by Wolfenstein (1949, 1954), Wolfenstein and Ashkin (1952), Dalitz (1952), Stapp (1955), Oehme (1955), Brown and Kanellopoulos (1957) and others. There is a valuable review article by Wolfenstein (1956 b).

For details of experiments and references, see the reviews by Fowler and Brolley (1956), Hulthén and Sugawara (1957), Taylor (1957) and Hess (1958), the book by Mather and Swan (1958) and the published proceedings of recent international conferences on high-energy nuclear physics at Rochester, 1955-6-7, (Interscience Publishers, New York) and CERN, 1958 (CERN Scientific Information Service, Geneva).

Charge independence. p–p scattering occurs in $T = 1$ states only, n–p scattering in both $T = 0$ and $T = 1$ states. Assuming charge independence, the pure $T = 0$ scattering can be deduced from n–p and p–p data, at angles where Coulomb effects are negligible. Thus Wolfenstein (1956 a) is able to deduce the $T = 0$ cross section and polarization near 300 MeV.

Charge independence, with the statement that n–p scattering contains an (unknown) $T = 0$ component, gives several inequalities between the cross sections $\sigma_{np}(\theta)$ and $\sigma_{pp}(\theta)$ (Jacobsohn 1953, Feldman 1953). For example

$$\sigma_{np}(\theta) + \sigma_{np}(\pi - \theta) - \tfrac{1}{2}\sigma_{pp}(\theta) \geqslant 0 \qquad \ldots\ldots(17)$$

(centre-of-mass angles), which states in effect that the pure $T = 0$ cross section cannot be negative. The high-energy data fulfil these inequalities (where Coulomb effects can be ignored).

To prove charge independence rigorously, we need to show that the n–p and p–p data (strictly we should include n–n too) *independently* define the same $\mathscr{S}$-matrix for $T = 1$ states; this requires many more data than are available. However, we may test whether n–p and p–p data can together be explained by a single set of phase shifts; this has presented no difficulty yet, perhaps because of the paucity of data (see § 2.6).

There is evidence for the weaker 'charge symmetry' hypothesis (invariance under n$\longleftrightarrow$p) in the similarity between the scattering of polarized neutrons from protons at 350 MeV (Siegel *et al.* 1956) and that of polarized protons from neutrons at 310 MeV (Chamberlain *et al.* 1957; this experiment is strictly a quasi-elastic

scattering from neutrons bound in deuterium). The unpolarized p–p and n–n cross sections also appear to agree (e.g. Hess 1958).

In short, present high-energy scattering data seem consistent with charge independence.

Some features of the scattering. $\sigma_{np}(\theta)$ and the nuclear part of $\sigma_{pp}(\theta)$ are isotropic in the low-energy S-wave region (< 10 Mev), but we expect anisotropy at higher energies.

This anisotropy quickly appears in $\sigma_{np}(\theta)$, which develops forward and backward peaks and appears roughly symmetrical about $\theta = 90°$, at least up to 90 Mev. To explain this a 'Serber potential' $(1+P_x)\,V(r)$ was proposed, giving no odd-parity interaction and making the cross section strictly symmetrical; the apparent absence of P-wave effects in early p–p experiments seemed to confirm it. However, later experiments destroy this simple model. The $\sigma_{np}(\theta)$ symmetry is not exact at 90 Mev, and becomes much worse at higher energies; p–p polarization comes entirely from odd-parity states; other n–p and p–p experiments reveal substantial odd-parity scattering.

Virtually no anisotropy appears in $\sigma_{pp}(\theta)$. The observed cross section (symmetrical about $\theta = 90°$ because the particles are identical) contains a forward Coulomb peak, followed by a region where Coulomb–nuclear interference dominates. At larger angles, where Coulomb effects are small (at high energies this is most of the angular range), it remains nearly isotropic until well beyond 300 Mev. None of the simple central potentials (3) can explain such behaviour.

Several potential models were proposed to explain these cross sections. Though not wholly satisfactory and now superseded, these models have historical importance and are described below. The attempt of Gammel *et al.* (1957) to fit $P(\theta)$ as well as $\sigma(\theta)$ is included here. The more successful potential of Gammel and Thaler (1957) is discussed separately in § 2.8.

The high-energy cross sections are now supplemented at various energies by the polarization parameter $P(\theta)$ and some triple-scattering and correlation parameters. These data are too numerous for a direct approach through potential models to have much chance of success. It is better first to make a phase-shift analysis (§ 2.6). Even if the phase shifts are known at only one energy, they may give enough insight for a successful return to a potential model (§§ 2.6, 2.8).

The Christian–Hart analysis. Christian and Hart (1950) analyse n–p cross sections at 40 and 90 Mev with static potentials which fit the low-energy data. Assuming that the potential has roughly the same shape in all states (and that this shape is fairly conventional), they conclude it is probably long-tailed like the Yukawa and exponential wells, and that it has roughly the Serber $1+P_x$ exchange character.

The Christian–Noyes potential. Christian and Noyes (1950) consider p–p cross sections at 32 and 340 Mev. Since the n–p potentials of the Christian–Hart analysis do not fit these data, they propose to abandon charge independence. They attribute the isotropy to a tensor force which is highly singular at $r = 0$ (or at least very strong for $r < 0{\cdot}5 \times 10^{-13}$ cm), and suggest a triplet potential $S_{12}V_t \exp(-\lambda r)/(\lambda r)^3$, where $V_t = \pm 18$ Mev and $\lambda^{-1} = 1{\cdot}6 \times 10^{-13}$ cm, plus a singlet square well of depth 13·3 Mev and range $2{\cdot}6 \times 10^{-13}$ cm (fitted to the low-energy data).

Swanson (1953) questions the use, by Christian and Noyes, of Born approximation for this singular potential and suggests that the tensor well be replaced at short distances by a square well or some combination of square wells. Using numerical integration, he chooses the modified tensor well

$$V(r) = V_t \exp(-\lambda r)/(\lambda r)^2 \quad \text{for} \quad r > r_1,$$

$$V(r) = V(r_1) \quad \text{for} \quad r_0 < r < r_1$$

and

$$V(r) = +\infty \quad \text{for} \quad r < r_0,$$

where $V_t = +15{\cdot}2$ MeV, λ is as before, $r_1 = 0{\cdot}48 \times 10^{-13}$ cm, and $r_0 = 0{\cdot}24 \times 10^{-13}$ cm.

The Case–Pais potential. Case and Pais (1950) ascribe the isotropy in $\sigma_{pp}(\theta)$ to an $\mathbf{L}.\mathbf{S}$ spin–orbit potential with a strong singularity at $r = 0$, and suggest the triplet potential

$$\mathbf{L}.\mathbf{S}\, V_t(\lambda_t r)^{-1} \frac{d}{d(\lambda_t r)} [(\lambda_t r)^{-1} \exp(-\lambda_t r)],$$

where $V_t \simeq \pm 24$ MeV and $\lambda_t^{-1} \simeq 1{\cdot}1 \times 10^{-13}$ cm, with the singlet potential

$$V_s(\lambda_s r)^{-1} \exp(-\lambda_s r)$$

where $V_s = -45{\cdot}8$ MeV and $\lambda_s^{-1} = 1{\cdot}18 \times 10^{-13}$ cm (fitted to the low-energy data). They remark that with such a $T = 1$ potential it may be possible to retain charge independence, and that the spin–orbit part has the same order of magnitude as that suggested by the shell model (§ 2.4.2).

The Jastrow potential. Jastrow (1951) shows that a strong short-range repulsion, at least in the singlet state, can explain the isotropy in $\sigma_{pp}(\theta)$ without violating charge independence. He takes the singlet potential $\frac{1}{2}(1 + P_x) V_s \exp[-\lambda_s(r - r_0)]$ cut off by a hard core at $r = r_0$, with $V_s = -375$ MeV, $\lambda_s^{-1} = 0{\cdot}4 \times 10^{-13}$ cm and $r_0 = 0{\cdot}6 \times 10^{-13}$ cm. The triplet potential is

$$[\tfrac{1}{2}(1 + P_x) + 1{\cdot}84(0{\cdot}3 + 0{\cdot}7P_x)\, S_{12}]\, V_t \exp(-\lambda_t r)$$

with $V_t = -69$ MeV and $\lambda_t^{-1} = 0{\cdot}75 \times 10^{-13}$ cm. This secures a rough fit to both n–p and p–p cross sections.

Goldfarb and Feldman (1952) recalculate the p–p scattering predicted by the Christian–Noyes, Case–Pais and Jastrow potentials, using numerical and variational methods. They put $V = 0$ at small distances in the tensor and $\mathbf{L}.\mathbf{S}$ potentials to remove the objectionable r^{-2} and r^{-3} singularities, and adjust some of the parameters slightly. The original (Born approximation) calculations for the singular potentials prove to be misleading, since these potentials really give quite large anisotropies in $\sigma_{pp}(\theta)$. The Jastrow potential continues to agree qualitatively with experiment.

The Gammel–Christian–Thaler analysis. These authors (Gammel *et al.* 1957) try to explain both $\sigma(\theta)$ and $P(\theta)$ (n–p and p–p) by a charge-independent static potential with hard cores. Yukawa shape is chosen, and the core radii are supposed not to depend on parity. There remain fourteen free parameters: the central and tensor depths and ranges for each spin and parity, and the core radii for each spin. These are adjusted by an electronic computer to fit the low-energy data, and $\sigma(\theta)$ and $P(\theta)$ up to 300 MeV. There is only limited success, however; the best overall fit makes $P_{pp}(\theta)$ much too small and even of the wrong sign.

2.4.5. *Other experiments.*

Radiative n–p capture, $n + p \rightarrow \gamma + d$. This process is negligible compared with elastic scattering except at very low (thermal) energies. The initial two-nucleon state is thus an S-wave, and we have to consider magnetic dipole (M1) transitions $^1S_0 \rightarrow {}^3S_1$ and electric quadrupole (E2) transitions $^3S_1 \rightarrow {}^3D_1$; the latter prove to be negligible.

We require the initial and final wave functions in order to make calculations. In the limit of zero range forces, each wave function is given by its asymptotic form, fixed by the phase shift (initial state) and deuteron binding energy (final state). Bethe and Longmire (1950) derive approximate range corrections in terms of the n–p effective ranges. Corrections for potential shape are not large, for plausible potential models (including hard-core models). To this approximation, ignoring interaction currents and relativistic corrections, the cross section is fixed by the low-energy parameters. There is a small discrepancy here between theory and experiment; it seems arguable whether or not this is covered by experimental and other uncertainties.

A discussion of the theory is given by Hulthén and Sugawara (1957). Baker and Wilkinson (1958) give the latest experimental values.

Deuteron photodisintegration, $\gamma + d \rightarrow n + p$. Very close to the threshold energy, this is simply the inverse of the (mainly M1) low-energy capture process above. However, the E1 transitions $^3S + {}^3D \rightarrow {}^3P$ soon become important; the E1 cross section is comparable with the M1 cross section at 2·5 MeV photon energy, and becomes much larger thereafter.

If the final P-state is represented by a free wave function (neglecting both phase shift and short-range distortion by nuclear forces), the deuteron and final S-state are treated as in the capture process, and interaction currents are ignored, the low-energy photodisintegration cross section is essentially fixed by the low-energy n–p parameters. Theory and experiment agree well on the total cross section, though less well on the angular distribution.

At higher energies a potential model is needed to give the wave functions, and higher multipole transitions may have to be included. The high-energy cross sections differ widely from the predictions of early (mainly Serber-exchange) n–p potentials, however. This was taken to show that the simple interaction mechanism assumed (ignoring interaction currents) was wrong; strong mesonic effects were proposed (Wilson 1956), with a violation of Siegert's theorem (Austern 1957).

For accounts of theory and experiment up to this point, see the articles by Hulthén and Sugawara (1957), Bishop and Wilson (1957) and Wattenberg (1957).

Recently, more sophisticated potential models have reached fair agreement with experiment at moderate energies without invoking interaction currents. Calculations are reported for a modified version of the Signell–Marshak potential (in particular, the spin–orbit term is omitted in $T = 0$ states, see §§ 2.4.6, 3.3.10) at 22, 54 and 80 MeV photon energies (de Swart and Marshak 1958, see also Zernik *et al.* 1959), and for the Gammel–Thaler potential at 120 MeV (Nicholson and Brown 1959). The agreement depends partly on a substantial deuteron D-state probability, $P_D \sim 7\%$ (cf. § 2.4.1).

Deuteron photodisintegration is also suggested as a source of information about high-energy nucleon–nucleon phase shifts, to supplement scattering data (Hsieh and Nakagawa 1956, Hsieh 1956 a, b, 1957, 1958). A reasonable deuteron wave function is assumed, the final wave functions are determined by phase shifts (neglecting short-range distortion) and appropriate multipole transitions are calculated (ignoring interaction currents).

On the other hand, we can describe deuteron photodisintegration (and the inverse process) empirically in terms of the $\mathscr{S}$-matrix elements for this reaction, making no assumptions about nuclear forces or the deuteron's electromagnetic structure (Simon 1953, Morita *et al.* 1954, Kawaguchi 1958). These matrix elements can be related by unitarity arguments to those for n–p and γ–d scattering, but there are not enough data at present to deduce any of the former.

Only the cross section has been measured up to the present, but other measurements are conceivable. Polarization effects have been investigated theoretically, both for photodisintegration and the inverse process (Czyż and Sawicki 1957 a, b, Kawaguchi 1958, de Swart *et al.* 1959, Zernik *et al.* 1959).

To summarize, however, though it seems possible to explain the data under certain assumptions, deuteron photodisintegration is not at present a very reliable guide to nuclear forces.

Electron–deuteron scattering. At high energies this depends on the electromagnetic structure of the deuteron. If we assume the nucleons can be treated as point charges and point dipoles, and ignore interaction currents, this structure is determined directly by the wave function and thus by the n–p potential. Jankus (1956) gives formulae and sample calculations for elastic and inelastic scattering.

It is perhaps more reasonable to include a proton charge distribution, as given by e–p scattering (low-energy experiments suggest the neutron charge distribution is negligible : see the review by Foldy (1958)). The deuteron electromagnetic structure still depends directly on the wave function. The e–d data are consistent with this model, but do not at present distinguish between Yukawa and Gartenhaus (repulsive core : § 3.3.6) n–p potentials (McIntyre 1956, McIntyre and Dhar 1957).

It is not clear how serious the neglect of relativistic and interaction-current effects may be.

For further references and discussion, see the papers cited† and the reviews by Yennie *et al.* (1957), Hofstadter (1957) and Hofstadter *et al.* (1958).

Nucleon–deuteron scattering. This would afford useful information about nuclear forces, in particular the n–n force, were not the three-body problem so intractable.

For an account of theory and experiment at low energies, see the review by Massey (1953); for more recent work see, for example, Bransden *et al.* (1958) and references therein. No fresh information on the two-nucleon force has yet come from this source.

At high energies, using the impulse approximation (see Chew and Wick 1952) the nucleon–deuteron scattering matrix can be related to the nucleon–nucleon scattering matrix‡ (strictly, we need to extrapolate the latter off the energy shell).

† See also recent work by Blankenbecler (1958), Goldberg (1958) and McIntyre and Burleson (1958).

‡ A similar approach has been used to explain p–α polarization, using the Gammel–Thaler two-nucleon potential (McManus *et al.* 1958).

This appears to offer fresh information. However, the different sets of phase shifts which fit n–p and p–p data all account tolerably well for the present nucleon–deuteron data in this approximation (L. Castillejo and L. S. Singh 1959, Sawicki and Watanabe 1959).

Nucleon–nucleus scattering (complex nuclei). At high energies the nucleon–nucleus scattering matrix can be related approximately to the free nucleon–nucleon scattering matrix (Takeda and Watson 1954, 1955, Riesenfeld and Watson 1956, Brown 1957, Bethe 1958, McCauley and Brown 1958).

The nucleon–nucleus cross sections and polarizations agree roughly with the Gammel–Thaler and Signell–Marshak potentials (§ 2.8, 3.3.10) (McManus and Thaler 1958, Ohnuma 1958), and are not inconsistent with any of the five 'best' phase-shift solutions for p–p scattering at 310 MeV (Bethe 1958; § 2.6). The Feshbach–Lomon model (§ 2.5) also appeared to fit the data at first (Riesenfeld and Watson 1956), but this was due to a particular choice of scattering angle; more careful examination reveals substantial disagreement (Ohnuma 1958, and private communication from R. M. Thaler).

The reaction $p+p\rightarrow\pi^{+}+d$. This reaction can be described semi-empirically at energies not far above threshold, assuming the pion to be emitted in an S- or P-state; if π–d scattering is ignored, the parameters of this description can be related by unitarity arguments to the $^1S_0$, $^3P_1$ and $^1D_2$ p–p phase shifts at the incident energy (Gell-Mann and Watson 1954). These parameters have all been measured (Tripp 1956); they are used to eliminate certain solutions to the 310 MeV p–p phase shift analysis of Stapp *et al.* (1957).

2.4.6. *Spin–orbit coupling.*

Spin–orbit terms $V(r)\,\mathbf{L}.\mathbf{S}$ in the two-nucleon potential have been proposed in recent years, and have proved very useful. The evidence for them we summarize below.

The shell model of the nucleus suggests such a term (see § 2.4.2), but first-order estimates disagree on the isobaric-spin dependence.

We may also expect such terms on general grounds, as relativistic Thomas corrections. Several authors observe that a repulsive core can give a large short-range correction, of the sign suggested by high-energy p–p scattering.

Wolfenstein (1956 c) analyses the 310 MeV p–p data near $\theta = 90°$ in terms of the amplitudes $A, B, \ldots$, etc. of (5), and finds a substantial B term which cannot come from a central or tensor potential in Born approximation, though it can come from an $\mathbf{L}.\mathbf{S}$ spin–orbit potential. The 310 MeV p–p phase-shift analysis of Stapp *et al.* (1957) has five plausible solutions; again, none of these are consistent—to a first-order approximation—with a central plus tensor potential. Born approximation can be misleading, but the investigation of Gammel *et al.* (1957, see § 2.4.4) with specific potential models supports the conclusion that central and tensor forces are not enough, at least in $T = 1$ states. Accepting the need for some non-static terms, it is natural to consider first those linear in $\mathbf{p}$, of which the spin–orbit potential is the only acceptable form (§ 2.2). An attractive $T = 1$ spin–orbit term has great success (§§ 2.8, 3.3.10).

The need for a $T = 0$ spin–orbit term is less well established. Gammel and Thaler (1957) find that such a term improves the fit to $\sigma_{np}(\theta)$ near 100 MeV, when

added to a central plus tensor potential *previously fitted* to the low-energy data, but this improvement disappears when the new potential is readjusted slightly to re-fit the low-energy data (J. L. Gammel, private communication). de Swart *et al.* (1957) also remark that the $T = 0$ spin–orbit coupling in the Signell–Marshak potential (§ 3.3.10) is not indispensable.

Some Japanese physicists suggest that little or no spin–orbit coupling may be needed to fit the data up to 150 MeV, but no definite conclusion can yet be reached (Otsuki 1958, Tamagaki 1958, Watari 1958).

2.4.7. *Separable potentials.*

Non-local separable potentials, such that $(\mathbf{p}'|V|\mathbf{p}'') = \lambda g^*(\mathbf{p}')g(\mathbf{p}'')$, have mathematical advantages, but relatively little work has been done with them.

Yamaguchi (1954) considers a central interaction $g(\mathbf{p}) = g(p)$, which acts in S-states only. With the assumed form $g(p) = 1/(p^2+\beta^2)$, where β is a constant parameter, the low-energy n–p data (singlet or triplet) are fitted.

Yamaguchi and Yamaguchi (1954) discuss the triplet n–p interaction in terms of the non-central potential,

$$g(\mathbf{p}) = C(p)+(1/\sqrt{8})\{(3/p^2)(\boldsymbol{\sigma}^{(1)}.\mathbf{p})(\boldsymbol{\sigma}^{(2)}.\mathbf{p})-\boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}\}\,T(p) \qquad \ldots\ldots(18)$$

which acts only in the α-eigenstate (which tends to $^3S_1$ as $\epsilon^1 \to 0$) of the coupled $^3S_1$–$^3D_1$ states. Assuming the forms $C(p) = 1/(p^2+\beta^2)$ and $T(p) = -tp^2/(p^2+\gamma^2)^2$, where β, γ and t are constant parameters, they fit the triplet low-energy data. Deuteron photodisintegration is discussed with the help of the previously determined (central) singlet potential; there is less discrepancy with the high-energy cross section than is given by other potential models of that time.

The determination of a separable potential from phase shifts is considered by Gourdin and Martin (1957 a, b, 1958) and Chadan (1957, 1958 a, b); see § 2.7.

Several authors use separable potentials, for their special convenience, in studying the nuclear many-body problem (e.g. Martin and de Dominicis 1957, Sundaresan 1957).

Lévy (1952 b) shows that, in a certain approximation appropriate for small distances, the potential derived from meson theory contains a central separable term (see § 3.3.3).

2.5. *The Boundary Condition Model*

Meson theory (§ 3) suggests that the two-nucleon potential becomes non-local and very strong at small distances. One may try to represent this strong non-local core by imposing a boundary value for the logarithmic derivative of the wave function, with respect to r, at some small radius. For coupled states $L = J \pm 1$, the boundary condition is expressed by a symmetric 2×2 matrix. A simple local potential may be added outside the core. Both core radius and boundary value may depend on the state, but are expected to depend little on energy since the core potential should be much larger than bombarding energies of interest.

This model, with no outer potential, was originally proposed for low-energy S-wave scattering by Breit and Bouricius (1949); the core radius and boundary

value effectively replace a and r_0 (§ 2.4.3). The effects of any real outer potential are not wholly ignored, for a zero-range potential immediately outside the core can be incorporated simply by modifying the boundary value ; in particular, any boundary condition can be regarded as a hard core plus a zero-range potential (Breit and Hull 1953).†

Strictly speaking, Jastrow's (1951, § 2.4.4) singlet potential and subsequent hard-core potentials can be accommodated in the model, for a hard core gives an energy-independent boundary condition. However, the idea behind the model is that the outer potential should be relatively unimportant.

Feshbach and Lomon (1956) apply the model to high-energy scattering, again with no outer potential. Using partial waves with $L \leqslant 3$, they secure an approximate fit (with two variants) to available n–p and p–p cross-section and polarization data up to 270 MeV, and to the low-energy data. p–p scattering is ascribed mainly to $^1S_0$ and $^3P_0$ waves (which, alone, would give an isotropic cross section) ; the $^3P_0$ scattering is close to that from a hard sphere, with a large negative phase shift. All boundary parameters are energy independent, except the $^1S_0$ core radius. The core radii lie between 0·88 and $1{\cdot}32 \times 10^{-13}$ cm ; the latter value is larger than meson theory suggests for a strong core.

When there is no outer potential, the boundary parameters are simply another representation of phase shifts. If the former depend little on energy, there are restrictions of the behaviour of the latter with energy. Hence if the Feshbach–Lomon model were verified, it would imply restrictions on the shape of any (alternative) potential model which is to fit the data ; Raphael (1956) investigates this point.

However, Feshbach and Lomon do not try to fit Coulomb interference effects, nor n–p polarization, nor the angular distribution of p–p polarization. Saperstein and Durand (1956) observe that in these and other respects the fit to data is poor‡ ; they are unable to secure a good fit by making small adjustments. Saperstein and Durand then derive an independent fit to p–p data alone (including triple-scattering) using partial waves with $L \leqslant 4$, keeping core radii fixed but allowing boundary values to vary with energy. A small attractive singlet outer potential is introduced to improve the low-energy fit. There is strong $^1S_0$ and $^3P_0$ scattering at high energies. A tolerable fit to data is secured, but the phase shifts at 310 MeV are unlike any of the eight best solutions to the phase-shift analysis of Stapp *et al.* (1957 ; § 2.6) and moreover—like solutions 5, 7 and 8 of that analysis—disagree with the $p+p \rightarrow \pi^+ + d$ experiments (Tripp 1956, see § 2.4.5).

Solution 6 of the analysis of Stapp *et al.* (1957) corresponds roughly to the original Feshbach–Lomon model. Recently, Lomon (1958, and private communication) has fitted the $^1S_0$ phase shift at 310 MeV according to this solution, and at low energies, using energy-independent boundary parameters and a short-range attractive outer potential. The core radius is $0{\cdot}8 \times 10^{-13}$ cm, more reasonable than the Feshbach–Lomon value $1{\cdot}32 \times 10^{-13}$ cm.

† It is interesting that a hard core can sometimes be replaced for practical purposes by a boundary condition. Moszkowski and Scott (1958) show for example that the S-wave scattering from a certain square well plus hard core potential can be closely approximated, up to 300 MeV, by that from a suitable square well plus boundary condition potential. Higher-wave scatterings are not similar, however.

‡ There also appears to be indirect conflict with nucleon–nucleus scattering data (Ohnuma 1958 ; see § 2.4.5).

If this solution 6 were eliminated by further experiments,† or if an accurate phase-shift analysis were possible at some intermediate energy, the boundary condition model might need drastic modifications. It would lose its original character if large energy variations were required, or the outer potential became more important than the boundary condition.

2.6. *Phase-shift Analysis*

It is helpful to analyse scattering data at any particular energy in terms of phase shifts whenever possible. The latter are not yet well enough known for a potential to be constructed directly from them (§ 2.7), but even a limited knowledge can be useful.

Born approximation. For a central potential $V(r)$, the Born phase shifts are

$$\delta_L \simeq -\tfrac{1}{2}\pi M \int_0^\infty V(r)\,[J_{L+1/2}(kr)]^2\, r\, dr, \qquad \ldots\ldots(19)$$

for small δ_L (Mott and Massey 1949). $J_{L+1/2}(x)$ is the usual Bessel function of half-odd-integral order (Watson 1944); other quantities are defined in §§ 2.2, 2.3. In this integral $[J_{L+1/2}(kr)]^2$ acts as a weighting factor concentrated roughly speaking near the classical impact parameter.

Thus $\delta_L > 0$ (or < 0) means that the Lth partial wave experiences a net attraction (or repulsion). If, for example, δ_L were positive at low energies and negative at high energies, it would suggest a potential attractive at long range and strongly repulsive at short range (though more complicated explanations would be possible).

Non-central potentials distinguish between the triplet states of common L. There are expressions analogous to (19) for the phase shifts $\delta_L{}^J$ in which $V(r)$ is replaced by the diagonal matrix element of V for the particular spin state. For coupled states, this gives the 'bar' phase shifts $\bar{\delta}_L{}^J$ defined in § 2.3 (see Stapp *et al.* 1957). We may hope to identify tensor and spin–orbit potentials from the splitting (i.e. the ratios of differences) of a set of triplet phase shifts, although of course any two independent types of non-central potential can account for a given splitting.

Unfortunately, Born approximation is notoriously inaccurate for many potentials (Jost and Pais 1951, Kohn 1954), and provides at most only a qualitative basis for interpreting phase shifts. It fails altogether when $V(r)$ contains an infinitely repulsive ('hard') core, since (19) becomes infinite.

For hard core potentials, however, we can modify the usual derivation of (19) (Mott and Massey 1949) to include the core exactly and treat only the potential outside it as perturbation. The *additional* phase shift due to the outer potential is then given (to first order in this potential) by a formula like (19), where $[J_{L+1/2}(kr)]^2$ is replaced by $2kr/\pi$ times the square of the appropriate hard-core radial wave function, and the integral extends from the core radius to infinity. There is an analogous treatment for non-central outer potentials: the relative splitting of the triplet phase shifts is as before.

† In fact a p–p correlation experiment at Liverpool now appears to eliminate this solution (Ashmore *et al.* 1959); it is also excluded by a modified phase-shift analysis (see § 2.6).

With such first-order approximations, we may hope to discover some qualitative features of the potential.

Methods of analysis. When only one partial wave is scattered, the phase shift is given directly by the cross section (to within a sign). In general, however, the analysis is more difficult.

Experimental data are not exact. An exact phase-shift solution can exist only in the sense that least-squares fits to experiments, with assumed parametric forms, are reproduced exactly (only possible when the analysis is not overdetermined).† Unfortunately, such solutions are usually not unique.‡ For this reason, and also to check the underlying assumptions (only certain partial waves are scattered, charge independence, etc.), it is desirable to have enough data at the energy in question to overdetermine the phase shifts in principle.

Approximate solutions are then sought. The most powerful method is a simultaneous least-squares fitting of all the data directly in terms of phase shifts, using a high-speed electronic computer. If O_i $(i = 1, 2, \ldots)$ are experimental measurements—various quantities at various angles—with standard deviations Δ_i, and $O_i(\delta)$ are their predicted values for given phase shifts δ, the computer is set to minimize $\mathscr{M}(\delta) = \Sigma_i[O_i - O_i(\delta)]^2/\Delta_i^2$ by adjusting the set δ. The minimizing may be done in several ways (Stapp 1955, Stapp *et al.* 1957), for example by adjusting one phase shift at a time, or computing the gradient of $\mathscr{M}$ in phase-shift space and moving down it, or taking random steps, or some combination of these. Many random starting-points are chosen, for clearly the minimum found depends partly on the starting-point. $\mathscr{M}(\delta)$ at the minimum measures how well the data are fitted. This objective measure is valuable, but there are dangers in taking it wholly at face value ; it is founded on the assumptions that experimental errors are purely statistical§ and that only certain partial waves are scattered.

If the available nucleon–nucleon data admit several plausible solutions at a given energy, we may be able to reduce the number by other means. If solutions—or better, a unique solution—are known at a neighbouring energy, the continuity of phase shifts with energy may be invoked. On the other hand, we may try to relate the phase shifts to the reaction $p + p \rightarrow \pi^+ + d$, to nucleon–deuteron and nucleon–nucleus scattering, and to deuteron photodisintegration (§ 2.4.5). These relations depend upon various assumptions and approximations which may not be good, but treated with proper caution they are a helpful supplement to the nucleon–nucleon data.

In this context it may be useful to note that the derivative of a phase shift with respect to energy has a well-defined lower bound if the interaction vanishes beyond a certain distance (Wigner 1955, Goebel *et al.* 1955). This comes fundamentally from the principle of causality.

While considering the analyses reported below, one should remember that the earlier work is based on less complete and generally less accurate data. The present high-energy data are still confined to relatively few energies, often widely separated.

† The expressions for observed quantities in terms of phase shifts are difficult to invert, in general. Clementel and Villi (1955) describe an ingenious mechanical analyser for the relatively simple case of p–p scattering, with S and P waves.

‡ Examples of non-uniqueness are given by Clementel *et al.* (1957), Zastavenko *et al.* (1958) and others.

§ Thus experiments may be misleading when the statistical errors are forced lower than possible systematic errors.

Analyses with S and P waves ($L \leqslant 1$). The high-energy p–p cross section is too large for pure S-wave scattering, but it was at first hoped to explain experiments up to 200 MeV or higher with S and P waves alone. For p–p scattering there are four phase shifts; for n–p scattering, two more.

Thaler and Bengston (1954) explain $\sigma_{pp}(\theta)$ at 240 MeV in this way, with considerable freedom in the choice of phase shifts. Their results are compatible with $\sigma_{np}(\theta)$ at 260 MeV if a ^3D phase shift (the same for all three states: no S–D coupling) is allowed, but not with later polarization data near 300 MeV (Fried 1954).

Garren (1953, 1954, 1955, 1956) analyses the p–p cross section and polarization near 200 MeV. Ignoring data from the Coulomb interference region, a continuum of solutions is found; when these data are included, four solutions remain.

C. A. Klein (1955) analyses p–p and n–p cross-section and polarization data at 170, 260 and 330 MeV, and finds solutions which vary smoothly with energy. This is not strictly an S- and P-wave fit, for he includes a ^3D phase shift (like Thaler and Bengston) at the two higher energies.

Clementel and others (Beretta *et al.* 1955 a, b, Clementel and Villi 1955, Clementel *et al.* 1957) analyse p–p cross sections and polarizations from 18 to 260 MeV. They eliminate some solutions by requiring continuity with energy; two solutions remain.

However, there is now evidence of higher waves at most of the energies considered above. $P_{np}(\theta)$ exhibits large ^3D-wave effects at 95 MeV (Stafford *et al.* 1957), $P_{pp}(\theta)$ shows appreciable ^3F-wave effects as low as 118 MeV (Wilson 1957), and very accurate measurements of $\sigma_{pp}(\theta)$ reveal ^3F-wave effects even at 40 MeV (Noyes and MacGregor 1958).

Analyses with $L \leqslant 2$. There are five phase shifts for p–p scattering, and six more for n–p scattering.

Clementel and Villi (1957) extend their p–p analysis to n–p cross sections from 27 to 215 MeV, assuming the ^3D phase shifts to be equal (no S–D coupling). They also try to fit both $\sigma_{np}(\theta)$ and $P_{np}(\theta)$ at 95 MeV, including S–D coupling and assuming $\delta_2{}^2 = \delta_2{}^3 = 0$, but find the data incompatible with their assumptions.

Phillips (1957 b) analyses the p–p and n–p cross sections and polarizations at 95 MeV, using all eleven phase shifts. Three solutions are found, two corresponding roughly to the Feshbach–Lomon model (§ 2.5) and the third to the Gammel–Thaler and Signell–Marshak potentials (§§ 2.8, 3.3.10). In principle the analysis is overdetermined, which appears to confirm charge independence, but experimental errors and the imperfections of the solutions weaken this conclusion. The data contain a strong hint of F-wave scattering, hoped to be small.

MacGregor (1958) reports analyses of $\sigma_{pp}(\theta)$ from 4 to 20 MeV. The data allow continuous families of solutions; arguments based on the second-order pion potential (§ 3.3.4) are given for reducing their number. At 40 MeV higher waves are needed for a good fit to $\sigma_{pp}(\theta)$ and a measurement of P_{pp} (45°) (Noyes and MacGregor 1958).

Analyses with $L \leqslant 3$. There are nine phase shifts for p–p scattering. n–p scattering involves seven more; there are not yet enough data at any energy to analyse the latter.

Thaler *et al.* (1954) continue the work of Thaler and Bengston (1954) by including $^3F_2$ waves and $^3P_2$–$^3F_2$ coupling; they are able to fit $\sigma_{pp}(\theta)$ and $\sigma_{np}(\theta)$ near 240 MeV with some freedom.

Stapp (1955) analyses p–p data at 300 MeV ($\sigma(\theta)$, $P(\theta)$ and two triple-scattering experiments) by least-squares fitting with a high-speed computer. Many reasonable solutions are found, but only one proves at all satisfactory when tested against data in the Coulomb–nuclear interference region (originally omitted) and against a third triple-scattering experiment. This solution also agrees with the $p + p \rightarrow \pi^+ + d$ experiments (Tripp 1956).

Hull *et al.* (1956) analyse p–p data at 300 MeV ($\sigma(\theta)$ and $P(\theta)$ only) by a different method. Using eight phase shifts (assuming $\epsilon^2 = 0$) they find several solutions, none wholly satisfactory, none corresponding to Stapp's best fit (above). Introducing some higher partial waves seems to be helpful.

Table 1

Sol'n	1	2	3	4	6
$\mathscr{M}$	17·9	21·7	23·8	24·5	34·6
δ_0	$-10{\cdot}1 \pm 2{\cdot}5$	$-19{\cdot}5 \pm 2{\cdot}0$	$-11{\cdot}0 \pm 2{\cdot}5$	$-27{\cdot}0 \pm 2{\cdot}0$	$-0{\cdot}3 \pm 2{\cdot}3$
δ_2	$12{\cdot}9 \pm 0{\cdot}7$	$4{\cdot}4 \pm 0{\cdot}6$	$13{\cdot}3 \pm 0{\cdot}8$	$4{\cdot}9 \pm 0{\cdot}6$	$12{\cdot}9 \pm 0{\cdot}6$
δ_4	1·0	1·3	1·1	1·1	−1·1
$\delta_1{}^0$	$-14{\cdot}3 \pm 2{\cdot}2$	$-36{\cdot}1 \pm 1{\cdot}9$	$-4{\cdot}1 \pm 1{\cdot}4$	$-25{\cdot}4 \pm 1{\cdot}9$	$-64{\cdot}7 \pm 1{\cdot}9$
$\delta_1{}^1$	$-26{\cdot}7 \pm 1{\cdot}3$	$-11{\cdot}7 \pm 1{\cdot}0$	$-19{\cdot}8 \pm 0{\cdot}8$	$-7{\cdot}3 \pm 1{\cdot}0$	$-13{\cdot}4 \pm 0{\cdot}9$
$\bar{\delta}_1{}^2$	$16{\cdot}1 \pm 1{\cdot}0$	$18{\cdot}8 \pm 1{\cdot}1$	$22{\cdot}6 \pm 0{\cdot}6$	$23{\cdot}1 \pm 1{\cdot}1$	$8{\cdot}2 \pm 0{\cdot}5$
$\bar{\epsilon}^2$	$-1{\cdot}0 \pm 0{\cdot}6$	$-9{\cdot}3 \pm 0{\cdot}9$	$1{\cdot}8 \pm 0{\cdot}4$	$-7{\cdot}5 \pm 0{\cdot}7$	$-0{\cdot}2 \pm 0{\cdot}6$
$\bar{\delta}_3{}^2$	$0{\cdot}8 \pm 1{\cdot}0$	$-0{\cdot}5 \pm 1{\cdot}2$	$-2{\cdot}0 \pm 0{\cdot}6$	$-1{\cdot}4 \pm 1{\cdot}2$	$-2{\cdot}1 \pm 0{\cdot}7$
$\delta_3{}^3$	$-4{\cdot}4 \pm 1{\cdot}1$	$0{\cdot}3 \pm 0{\cdot}4$	$-2{\cdot}6 \pm 0{\cdot}6$	$1{\cdot}6 \pm 0{\cdot}4$	$3{\cdot}1 \pm 1{\cdot}1$
$\delta_3{}^4$	$3{\cdot}2 \pm 0{\cdot}5$	$2{\cdot}5 \pm 0{\cdot}8$	$0{\cdot}5 \pm 0{\cdot}5$	$2{\cdot}6 \pm 0{\cdot}8$	$3{\cdot}3 \pm 0{\cdot}3$
$\bar{\epsilon}^4$	−1·2	−1·5	−0·9	−0·9	1·3
$\bar{\delta}_5{}^4$	1·5	2·1	−1·1	−0·7	2·2
$\delta_5{}^5$	0·1	−1·4	0·9	−0·9	−2·0
$\delta_5{}^6$	1·3	1·4	−0·6	−0·8	0·3

The five best solutions of Stapp *et al.* (1957) at 310 MeV. Quasi-nuclear phase shifts are given, in degrees: 'bar' parameters are used for the coupled states.

Ohnuma and Feldman (1956) analyse $\sigma_{pp}(\theta)$ and $P_{pp}(\theta)$ near 150 MeV. There are only enough data to fix one F-wave parameter, which they choose to be $\delta_3{}^2$; $\delta_3{}^3$, $\delta_3{}^4$ and ϵ^2 are assumed to vanish. Several different solutions are found.

Clementel *et al.* (1957) briefly consider the effect of $^1D_2$ and $^3F_2$ waves at 170 MeV as a perturbation on their S- and P-wave results (above).

Phillips (1957 a) treats $\sigma_{pp}(\theta)$ and $P_{pp}(\theta)$ at 140 MeV. Again the data only fix one F-wave parameter, taken to be $\delta_3{}^2$ or $\delta_3{}^4$. The solutions for the two cases are quite dissimilar, showing that F-waves are not a small perturbation at this energy. A proper analysis requires more data.

Noyes and MacGregor (1958) analyse $\sigma_{pp}(\theta)$ and P_{pp} (45°) at 40 MeV. The data definitely require F-waves for a good fit, but are inadequate to define precise solutions.

Analysis with $L \leqslant 5$. Stapp *et al.* (1957) analyse the 310 MeV p–p data ($\sigma(\theta)$, $P(\theta)$ and three triplet-scattering experiments) by least-squares fitting with all 14 phase shifts. Eight solutions are found (numbered 1 to 8 by the authors) within wide limits of statistical plausibility as measured by $\mathscr{M}$. For the 'true' solution, there is a 90% chance in this analysis that $14 \lesssim \mathscr{M} \lesssim 34$, a 5% chance that $\mathscr{M} > 34$ and a 1% chance that $\mathscr{M} > 40$; the rejected solutions all have $\mathscr{M} > 62$.

Of the best eight solutions, Nos. 5, 7 and 8 are incompatible with the $p+p \rightarrow \pi^+ + d$ experiments (Tripp 1956; § 2.4.5); the remaining five are shown in table 1. The errors quoted (for $L \leqslant 3$ only) are computed from an error matrix (see Anderson *et al.* 1955).

If we interpret phase shifts qualitatively by the methods discussed earlier, $\delta_0 < 0$ in all five solutions suggests a strong short-range repulsion in the singlet state. As for the triplet states, no solution corresponds to a simple central plus tensor potential; at least one other type of non-central force is required (the results of Gammel *et al.* (1957; § 2.4.4) support this conclusion).

Assuming a central plus tensor plus **L.S** potential, No. 1 is the most easily understood; Gammel and Thaler (1957) base their potential on it (§ 2.8). Nos. 2 and 4 appear consistent with zero tensor force, except for their large values of $\bar{\epsilon}^2$. The tensor force deduced from the ^3P part of No. 3 has the wrong sign to explain $\bar{\epsilon}^2$. It is believed that No. 6 would be very difficult to fit with a potential (Gammel and Thaler 1957).

Solutions 2 and 4 correspond quite closely to Stapp's best fit with $L \leqslant 3$ (4 is somewhat closer than 2). No. 6 corresponds roughly to the Feshbach–Lomon model (§ 2.5).

It is hoped to eliminate some solutions by further experiments. A p–p correlation experiment at 380 MeV (Ashmore *et al.* 1958) unfortunately excludes none.† An appeal to proton–nucleus scattering also proves unhelpful (Bethe 1958); it was once stated that No. 6 could be eliminated but this is now withdrawn.

General. Most analyses employ the relativistic corrections to the kinematics and to the Coulomb scattering mentioned in § 2.3. The magnetic moment corrections are somewhat unreliable, because of wave function distortion: when only p–p data are analysed, these corrections are often assumed to be absorbed in the quasi-nuclear phase shifts.

Several workers (Phillips 1957 b, and others, unpublished) remark that it is difficult to fit the small angle points in $\sigma_{pp}(\theta)$, on the slope of the Coulomb peak (allowing, of course, for the experimental resolution). This may be due to unexpectedly high partial waves, or some experimental effect, but nevertheless remains to be explained.

Most of the analyses depend on inadequate data and make questionable assumptions about which partial waves are important. Of them all, that of Stapp *et al.* (1957) is the most satisfactory. To gain a sound empirical knowledge of nucleon–nucleon scattering, we require more data and reliable analyses at many energies.

Moravcsik's modified method. Moravcsik (1958) suggests a new technique of analysis. While the scattering in states of lower angular momentum is described by empirical phase shifts as before, the scattering *in all higher states* is calculated from pion field theory (§ 3.3) and included in the analysis. The calculation takes account of one-pion exchange processes in Born approximation, and is valid for sufficiently high angular momenta at any given energy. It is hoped to widen the range of validity by including two-pion exchange, but this cannot yet be done reliably. An interesting feature is that the pion–nucleon coupling constant g can

† However, a different correlation experiment now appears to exclude solution 6 (Ashmore *et al.* 1959).

be treated as an unknown parameter, to be determined with the phase shifts. This method also requires less data in general, and may yield fewer solutions.

Cziffra *et al.* (1958) use this approach to the 310 MeV p–p data, starting from the five best solutions of Stapp *et al.* (1957; see above). They calculate the contribution from partial waves with $L>5$ (previously ignored) for various values of the pseudoscalar coupling constant g, and re-adjust the empirical phase shifts ($L\leqslant 5$) to minimize $\mathscr{M}(\delta)$. The goodness of fit to the data is then studied as a function of g. The modified solutions 1 and 2 both give their best fits near the currently accepted value $g^2/4\pi\sim 14$, which is very satisfactory. Moreover, near

Table 2

Solution	1	2
$\mathscr{M}$	14·0	17·6
δ_0	−10·9	−22·1
δ_2	12·1	4·4
δ_4	1·2	1·1
δ_1^0	−14·0	−34·4
δ_1^1	−26·2	−11·1
δ_1^2	16·6	19·2
$\bar{\varepsilon}^2$	− 1·4	− 8·6
$\bar{\delta}_3^3$	1·3	0·1
δ_3^3	− 4·6	− 0·5
$\bar{\delta}_3^4$	3·2	2·9
$\bar{\varepsilon}^4$	− 1·4	− 1·7
$\bar{\delta}_5^4$	1·7	2·4
δ_5^5	− 0·4	− 1·5
δ_5^6	1·2	1·4

The modified solutions of Cziffra *et al.* (1958) at 310 MeV; notation as for table 1.

this value solution 4 ceases to exist as a separate minimum of $\mathscr{M}(\delta)$ and merges with No. 2, while at somewhat larger values of g, No. 3 similarly merges with No. 1. We are thus led to regard Nos. 3 and 4 as numerical accidents rather than separate solutions. Of the remaining solutions, Nos. 1 and 2 give the best fits to data, but No. 6 is still acceptable.

However, it seems reasonable at 310 MeV to extend the calculated scattering down to $L=4$. When this is done, and the phase shifts for $L\leqslant 3$ are re-adjusted, the modified solutions 1 and 2 still fit the data well but No. 6 becomes unacceptable (private communication from M. J. Moravcsik). There are now also other grounds for rejecting No. 6; see above.

Only the modified solutions 1 and 2 remain; they are shown in table 2 for the value $g^2/4\pi=14{\cdot}4$ (treating $L\leqslant 5$ empirically).

2.7. *The Construction of Potentials from Phase Shifts*

Scattering experiments are the main source of information about the two-nucleon potential; it is natural to ask how far the latter is determined by the phase shifts. Clearly the potential cannot be determined uniquely, for infinitely many give the same scattering and bound-state energies (§ 2.2), but the possibilities can be reduced by making assumptions about its form.

Static central potential. This was the first case to be studied. Early work (Fröberg 1947, Hylleraas 1948) suggested $V(r)$ was completely determined by any single phase shift δ_L, given as a function of energy. However, Bargmann (1949 a) showed that if there were a bound state with angular momentum L, more than one potential could give the same phase shift and binding energy. More careful analysis (Levinson 1949, 1953, Bargmann 1949 b, Gel'fand and Levitan 1951 a, b, Jost and Kohn 1952 a, b, 1953, Holmberg 1952) led to the following result. If

$$\int_0^\infty r|V(r)|\,dr \quad \text{and} \quad \int_0^\infty r^2|V(r)|\,dr$$

converge, and there are no bound $L=0$ states, $V(r)$ is uniquely determined by the phase shift δ_0 at all energies; if there are m bound states, there is an m-parameter set of potentials $V(r)$ with the same binding energies and phase shift. There are analogous results for $L>0$. $V(r)$ is thus uniquely determined in principle by the phase shift for any L sufficiently large for there to be no bound state.

Of course, given phase shifts may not be consistent with a static central potential; the $V(r)$ determined by different δ_L may not conicide, or the construction procedure may fail.† On the other hand, we can only find nucleon–nucleon phase shifts experimentally up to a finite energy, and must extrapolate them arbitrarily beyond that point. This added freedom in practice does not seem to have been studied.‡

The most elegant method for constructing the manifold of 'equivalent potentials' corresponding to given binding energies and phase shift is due to Gel'fand and Levitan (1951 a, b) (used and discussed by Jost and Kohn 1953, Levinson 1953 and Regge 1958). $V(r)$ is shown to be uniquely determined by a certain 'spectral function', which incorporates the phase shift, the m binding energies E_i and m corresponding free real parameters $C_i \geqslant 0$. The construction of $V(r)$ depends on the solution of a Fredholm integral equation.

It emerges clearly that there is no general connection between binding energies and the $\mathscr{S}$-matrix; potentials can be constructed with the same phase shift but arbitrary binding energies. However, if V is such that

$$\lim_{r\to\infty} V(r)\exp 2Kr = 0,$$

the $\mathscr{S}$-matrix—considered as a function of k in the complex plane—is analytic in a strip of width K above the real axis, except for a number of poles on the imaginary axis which are in 1–1 correspondence with bound states. Thus, for example, the connection between the phase shift and deuteron binding energy in effective range theory (see § 2.4.3) is valid in general only for potentials which tend to 0 as $r\to\infty$ faster than $\exp(-2K_1 r)$, K_1^2/M being the binding energy (Jost and Kohn 1953).

There is a possible range criterion for making a unique choice from a set of equivalent potentials. If the free parameter C_i corresponding to the binding

† To determine the spectral function (see below) we must compute integrals involving the phase shift, which may not converge. Again, there may exist no potential corresponding to a given spectral function: Jost and Kohn (1953) give some sufficient conditions that one may exist.

‡ Raphael (1957) considers this indeterminacy from another point of view.

energy $E_i = K_i^2/M$ is changed, the change in $V(r)$ behaves asymptotically as $\exp(-2K_i r)$. Hence if there is a potential in the set which decays asymptotically faster than $\exp(-2K_{\max} r)$, where $K_{\max} = \max K_i$, it is unique. However, there may be no such potential; for example, if any one decays more slowly than $\exp(-2K_{\min} r)$, where $K_{\min} = \min K_i$, then all have this property (Newton 1956, Ohmura 1956).

As yet there are few physically interesting applications of these results. Jost and Kohn (1953) construct explicitly the manifold of potentials corresponding to a single bound state at arbitrary energy $E_1 = K_1^2/M$ and S phase shift

$$k \cot \delta_0 = -1/a + \tfrac{1}{2} r_0 k^2$$

(the shape-independent form extended to all energies). Newton (1957 a) observes that the manifold determined in this way by the low-energy triplet n–p data allows the deuteron charge form factor to be chosen anywhere within the widest limits (with implications for e–d scattering). However, the low-energy deuteron photodisintegration data introduce a restriction (on the electric dipole transitions†) and in fact select the 'unique' potential with asymptotic decay faster than $\exp(-2K_1 r)$. The result extends to a large class of manifolds, defined by a single bound state and a phase shift of special functional form (which can approximate any reasonable phase shift) and for which the effective-range approximations are valid at low energies: photodisintegration selects the 'unique' potential, when it exists. This provides an independent argument that the triplet n–p potential has a small range parameter (Newton 1957 b).

The potential determined by the n–p singlet shape-independent formula, extended to all energies, is given by Bargmann (1949 b), Jost and Kohn (1952 a) and Newton and Fulton (1957).

We may ask whether $V(r)$ can be constructed from a knowledge of all phase shifts at one energy, rather than one phase shift at all energies; this seems plausible classically. Verde (1957) gives an approximate prescription for very high energies; at lower energies the problem is very complicated.

Non-central potential. Newton and Jost extend the Gel'fand–Levitan method to the case of coupled waves $L = J \pm 1$ and a central plus tensor plus spin orbit potential‡ (Newton and Jost 1955, Newton 1955). There is a close analogy with the previous case, except that the $\mathscr{S}$-matrix must now satisfy certain conditions necessary for the existence of a short-range potential (in the sense that an appropriate number of moments

$$\int_0^\infty |V| r^n \, dr$$

converge). If it exists, the potential is uniquely determined by a spectral function which incorporates the $\mathscr{S}$-matrix elements for the two coupled waves, the m bound state energies and m real positive semi-definite matrices.

Newton and Fulton (Fulton and Newton 1956, Newton and Fulton 1957) construct explicit examples, and in particular a potential fitted to the low-energy

† Chadan (1955) observes that the slope of the disintegration cross section against energy at threshold, which depends on the magnetic dipole transitions, can help to select a unique potential. This slope has not yet been determined.

‡ Since the $\mathscr{S}$-matrix elements for the coupled waves contain three independent parameters, three independent types of potential are needed in general.

triplet n–p data. They assume the eigen-phase shift $\delta_0{}^1$ to obey the shape-independent formula exactly, $\delta_2{}^1$ to vanish always and ϵ^1 to have a simple energy dependence. The free parameters are chosen to fit the deuteron binding energy, quadrupole moment and D-state probability. Newton and Scofield (1958) re-adjust the parameters to allow for a spin–orbit interaction moment (see § 2.4.1). With the singlet potential mentioned above, this provides a phenomenological n–p potential which allows an explicit solution of the Schrödinger equation in terms of known functions.

Separable potentials (see §§ 2.2, 2.4.7). Gourdin and Martin (1957 a, b, 1958) show that a central separable potential, $g(\mathbf{p}) = g(p)$, can be determined from the S-phase shift, provided the latter fulfils certain conditions. If there is no bound state, the solution is unique. With one bound state there is a one-parameter set of solutions which fit the phase shift, but only one of these fits a given binding energy. Two or more bound S-states are inconsistent with a separable potential.

Gourdin and Martin construct various examples, including the potentials which give the shape-independent formulae exactly.

If a central potential contains a local part and a (non-local) separable part, the latter can be deduced from a knowledge of the former and the S-phase shift (Chadan 1958 a, b, also stated by Gourdin and Martin 1957 b).

Chadan (1957) considers the determination of a non-central separable potential of the form (18) which acts only in the α-eigenstate of the coupled $^3S_1$ and $^3D_1$ states.

Other work. Ohmura (1956) studies variational methods of obtaining approximate solutions, for static central and tensor potentials.

A more abstract approach to the determination of a potential from scattering data is made by Kay and Moses (1955, 1956, 1957).

2.8. *The Gammel–Thaler Potential*

Gammel and Thaler (1957) set out to fit all n–p and p–p data up to about 300 Mev with a charge-independent central plus tensor plus spin–orbit potential, assuming a hard core with a monotonic potential outside in each state.

p–p scattering, which is confined to singlet even and triplet odd states, is treated first. In the 310 Mev phase-shift analysis of Stapp *et al.* (1957, § 2.6), only solution 1 promises to fit this potential model; it is made the starting point. Assuming Yukawa shape for the singlet even outer potential, the adjustable parameters (range, depth and core radius) are fixed by the low-energy data and δ_0 at 310 Mev. δ_2 at 310 Mev proves insensitive to changes in shape when the other data are fitted; that it is near the right value is more a check on the potential model than on details of shape.†

In the triplet odd states this solution 1 shows central, tensor and spin–orbit effects‡ in the ^{3}P-waves, but little spin–orbit effect in the ^{3}F-waves. It suggests a long-ranged tensor force (attractive in the $^3P_0$ state) with a short-ranged spin–orbit force (repulsive in the $^3P_0$ state); an analysis at 18 Mev (Gammel and Thaler, unpublished) supports this conclusion. Yukawa shape is tentatively assumed for

† Although the phase-shifts are given fairly well it now appears that the singlet even potential may have to be non-static to secure really close agreement (Gammel and Thaler 1959).

‡ We can only identify these effects of course by first assuming a central plus tensor plus spin–orbit outer potential (treated in first order). It would be impossible if other non-central forces were also present.

the non-central outer potentials, though a modified shape (which tends to zero at the core) for the tensor part proves slightly better; we record below the Yukawa form. It is found possible to assume no outer central potential; the core itself provides a central force, of course.

The various parameters are adjusted to fit the data. In an obvious notation, writing the potentials as $V\exp(-\mu r)/\mu r$, these are

$$\left.\begin{array}{llll} {}^1V^+ & = -425{\cdot}5 \text{ Mev}, & {}^1\mu^+ & = 1{\cdot}45\times10^{13}\text{ cm}^{-1} \\ {}^3V_T{}^- & = \ \ 22 \text{ Mev}, & {}^3\mu_T{}^- & = 0{\cdot}8\ \times10^{13}\text{ cm}^{-1} \\ {}^3V_{LS}{}^- & = -7317{\cdot}5 \text{ Mev}, & {}^3\mu_{LS}{}^- & = 3{\cdot}7\ \times10^{13}\text{ cm}^{-1} \end{array}\right\}. \qquad (20)$$

The singlet even and triplet odd core radii are $0{\cdot}4$ and $0{\cdot}4125\times10^{-13}$ cm, respectively. The very large value of ${}^3V_{LS}{}^-$ has no physical significance; the spin–orbit potential has a very short range and is cut off by the core.

Singlet odd and triplet even potentials are added to fit the n–p data. These data are relatively few, and there is no precise phase-shift analysis to guide the choice of potential. As an expedient, certain Yukawa central plus tensor hard core potentials which fit the low-energy data (taken from the work of Gammel *et al.* 1957, § 2.4.4) are modified by the addition of a Yukawa spin–orbit term, with adjustable depth but the same short range as before. A combination is found which fits the 90 Mev data; it continues to fit the low-energy data approximately, except perhaps the deuteron magnetic moment (see Feshbach 1957 and § 2.4.1). However, to fit the n–p data at 156 and 310 Mev, the depths of the central and tensor potentials have to be varied somewhat. This is taken to show that their assumed Yukawa shape is wrong; it is hoped to find other shapes which will fit all the data without variation of parameters.† A search for such shapes has not yet been made, and we simply record the Yukawa parameters for each energy.

$$\left.\begin{array}{lll} {}^1V^- = \left\{\begin{array}{l}100\text{ Mev (at 0, 90, 156 Mev)}\\ 150\text{ Mev (at 310 Mev)}\end{array}\right\}, & {}^1\mu^- & = 1{\cdot}0\times10^{13}\text{ cm}^{-1} \\ {}^3V_C{}^+ = \left\{\begin{array}{l}-100{\cdot}7\text{ Mev (at 0, 90, 156 Mev)}\\ -60\text{ Mev (at 310 Mev)}\end{array}\right\}, & {}^3\mu_C{}^+ & = 1{\cdot}23\times10^{13}\text{ cm}^{-1} \\ {}^3V_T{}^+ = \left\{\begin{array}{l}-257\text{ Mev (at 0, 90 Mev)}\\ -230\text{ Mev (at 156 Mev)}\\ -175\text{ Mev (at 310 Mev)}\end{array}\right\}, & {}^3\mu_T{}^+ & = 1{\cdot}203\times10^{13}\text{ cm}^{-1} \\ {}^3V_{LS}{}^+ = -5000\text{ Mev} & {}^3\mu_{LS}{}^+ & = 3{\cdot}7\times10^{13}\text{ cm}^{-1} \end{array}\right\}$$

.....(21)

The singlet odd and triplet even core radii are $0{\cdot}5$ and $0{\cdot}4\times10^{-13}$ cm, respectively.

This potential is illustrated in figures 1 and 2. We show the $T=1$ potential (20) rather than the modified form in which the tensor shape is non-Yukawa, and the $T=0$ potential (21) appropriate to the range 0–90 Mev. It is qualitatively

† The physical potential may of course be genuinely velocity-dependent, such that no static potential can *exactly* replace it. However, we are concerned here with finding the best static potential, and may reasonably expect to do better with some other shapes.

similar to one proposed about the same time but on different grounds by Signell and Marshak (1957, 1958; § 3.3.10), shown in figures 12 and 13.

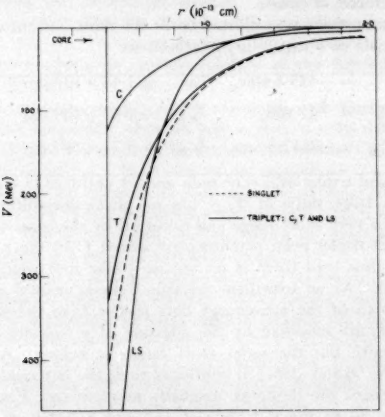


Figure 1. The Gammel–Thaler potential in even-parity states.

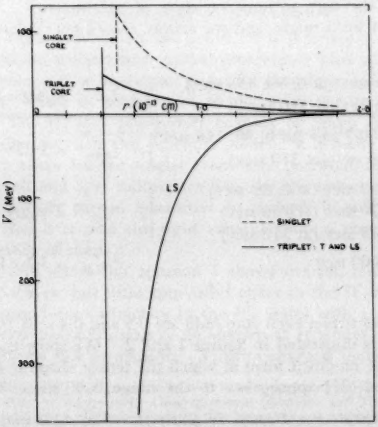


Figure 2. The Gammel–Thaler potential in odd-parity states.

The Gammel–Thaler potential agrees fairly well with all n–p and p–p data originally available up to about 300 MeV, with nucleon–nucleus scattering (McManus and Thaler 1958, Ohnuma 1958; § 2.4.5) and with deuteron photodisintegration at 120 MeV (Nicholson and Brown 1959; § 2.4.5). However, it disagrees with a later measurement of p–p 'depolarization'—a triple-scattering parameter—at 140 MeV at Harwell† (Taylor *et al.* 1958). The Signell–Marshak potential explains this quantity quite well but agrees poorly, on the other hand, with experiments above 150 MeV. The unequal successes of these potentials show that high-energy scattering is sensitive to details of the interaction.

2.9. *Outlook*

Two-nucleon phenomenology is left at present with no large anomalies to explain, but smaller problems remain. For example, it appears not inconsistent to ignore possible interaction currents in the deuteron (§§ 2.4.1, 2.4.5), but we are still far from an adequate empirical knowledge of them. It still seems arguable whether or not the thermal n–p capture cross section is accurately explained (§ 2.4.5). There is a difficulty in fitting the high-energy p–p cross section at small angles (§ 2.6). The agreement of the Gammel–Thaler potential with experiment is imperfect in several particulars (§ 2.8); small adjustments may be a sufficient remedy, but this is not certain.

There are other questions for the future. The Gammel–Thaler potential rests on the assumption that solution 1 to the p–p phase-shift analysis at 310 MeV (§ 2.6) is the correct one; if this is not confirmed by future experiments, a radically different potential will be indicated. Even if this solution is correct, experiments at other energies may show that the potential contains other non-static terms in addition to—or in place of—the $\mathbf{L}.\mathbf{S}$ term. The hypothesis of charge independence appears consistent but has yet to be confirmed rigorously at high energies.

Future advances will depend on more experiments, particularly high-energy nucleon–nucleon scattering. When precise phase-shift analyses can be made at several energies in the 40–300 MeV range, we shall have a much more accurate picture of the potential. The series of p–p triple-scattering experiments currently being performed at Harwell and Harvard will make possible an analysis at 140 MeV (with $L \leqslant 3$ or 4). It is hoped that the accurate work with proton linear accelerators will eventually be extended to polarization and correlation experiments (using polarized ion sources for the former), and thus make a precise analysis possible in the 50 MeV region. Moravcsik's new method (§ 2.6) now permits analyses at several energies where previously the data were inadequate. More data in the 200 MeV region—and n–p data at all energies—would be very welcome, however.

§ 3. Meson Field Theory

3.1. *Introduction*

This chapter is a survey of the ways in which a two-nucleon Schrödinger equation can be derived from a quantum field theory of nucleons interacting with

† This measurement, however, conflicts with preliminary results from Harvard, which agree with the Gammel–Thaler predictions (private communication from A. E. Taylor).

mesons, and of the results found in the physically interesting case of the pion (or π-meson) field.

In the absence of interactions, the Hamiltonian of a field theory has eigenstates describing the presence of any number of mesons, nucleons and nucleon–antinucleon pairs. When interactions are introduced, these states are coupled together in the equations of motion. The problem is to reduce this complicated and essentially relativistic scheme to a non-relativistic Schrödinger equation, which describes two nucleons interacting through an instantaneous potential.

The easiest and crudest simplification is to regard the two nucleons as fixed sources of the meson field, and to treat only the latter field-theoretically. The potential is defined as the potential energy of the two sources due to the meson field (subtracting the self-energies of isolated nucleons); the motion of the nucleons is calculated separately, using this potential in a Schrödinger equation. This is the 'adiabatic approximation'; it is quite successful in describing slow atomic collisions and diatomic molecules. However, one would prefer an approach which included nucleon recoil and which was field-theoretical throughout.

Several more sophisticated methods have therefore been devised, which are described in § 3.2. Incidentally, these methods can be adapted to the case of three or more nucleons.

We are concerned here only with essentially weak-coupling methods; these have received the greatest attention in recent years, and have had the most success. For strong-coupling methods†, the reader is referred to Pauli (1946) and Rosenfeld (1948); for intermediate coupling, to the *Progress of Theoretical Physics* Supplement No. 2 (1955) where many early papers are collected, Hasegawa (1955) and Nogami and Hasegawa (1956).

In § 3.3 we specialize to the pion field, and consider the two-nucleon potential to which it leads. In practice the adiabatic (fixed-source) potential is taken as a first approximation, and 'non-adiabatic' corrections are considered later. At short distances $\mu r \lesssim 0{\cdot}7$ (μ = pion mass, $\mu^{-1} = 1{\cdot}4 \times 10^{-13}$ cm), present methods of calculating the potential are quite inadequate, and it must be represented phenomenologically. At larger distances, the adiabatic potential is often approximated by the contributions from one- and two-meson exchange processes, ignoring radiative corrections. Even with these drastic approximations, there remain a variety of proposed potentials, corresponding to different choices of the pion–nucleon interaction and different treatments of certain controversial terms. These approximate adiabatic potentials are adjusted to fit the low-energy experimental data, but prove inadequate at high energies. When all possible corrections are considered, it appears that these potentials are reliable for $\mu r > 1{\cdot}5$ (where they agree) and perhaps qualitatively correct for $0{\cdot}7 < \mu r < 1{\cdot}5$.

Signell and Marshak are able to account fairly well for experimental data up to 150 Mev by adding an empirical spin–orbit coupling and a certain cut-off to one of the above potentials (the Gartenhaus potential). Pion field theory does predict a spin–orbit coupling among other non-adiabatic effects, but present methods do not allow a reliable estimate.

† In this context see also Matsumoto *et al.* (1953), Matsumoto and Sugawara (1954) and Zharkov (1958) for attempts to include nucleon isobars phenomenologically.

Nucleon–nucleon dispersion relations, in a static approximation, can be made to relate the adiabatic potential to observable amplitudes; however, there are not yet enough data for them to be useful. Assuming fixed nucleons, pion–nucleon dispersion relations can be used to infer the radiative corrections to two-pion exchange processes; they have been used by Miyazawa and others to calculate an approximate adiabatic potentia', and by Okubo to calculate spin–orbit corrections.

We ignore throughout the possible existence of a second neutral pion (discussed by Baldin (1958) and Yamaguchi (1958)). If such a light meson with strong nuclear interactions were to be discovered, the results reported in § 3.3 would need considerable modification.

As in § 2, we use units such that $\hbar = c = 1$.

3.2. *The Construction of a Potential*

3.2.1. *The canonical transformation method.*

This method uses canonical transformations to remove from the Hamiltonian terms that couple two-nucleon states to other states. When the two-nucleon subspace is decoupled, the equation of motion in it becomes a two-particle Schrödinger equation. In practice, the decoupling is performed only to a certain order in the coupling constant g. This approach to nuclear forces is due to Stückelberg (1938) and to Møller and Rosenfeld (1939, 1940); more recently it is used by Van Hove (1949), Nishijima (1951), Shindo and Nishijima (1955) and Iwadare (1955), and discussed by Feldman (1955).

Consider a nucleon field and a meson field in interaction, with the time-dependent Schrödinger equation

$$i\partial\Psi(t)/\partial t = H\Psi(t), \qquad \ldots\ldots(22)$$

where $\Psi(t)$ is the state functional in the Schrödinger representation at time t. $H \equiv H^0 + H'$ is the Hamiltonian, which is Hermitian; H^0 and H' are the free-field and interaction parts, respectively. H' contains a factor g. Let us use the eigenstates of H^0 to form the basis of a matrix representation, labelling the two-nucleon basic states by Greek letters $\alpha, \beta, \ldots$, and all others by Roman letters $j, k, \ldots$. The eigenvalues of H^0 we write E_α, E_j, etc. In general, H' has non-zero elements $H_{\alpha j}'$ coupling two-nucleon states to other states.

If we make the canonical transformation

$$\Psi(t) = \exp[iS'].\Psi'(t), \qquad \ldots\ldots(23)$$

where S' is Hermitian and independent of t, the Hamiltonian becomes

$$\exp[-iS']H\exp[iS'] = H + i[H, S'] - \tfrac{1}{2}[[H, S'], S'], \ldots, \text{etc.} \qquad \ldots\ldots(24)$$

Suppose S' contains the factor g; then in (24) the terms of zero and first order in g are

$$H^0 + H' + i[H^0, S']. \qquad \ldots\ldots(25)$$

The (αj) matrix elements vanish to this order if

$$S_{\alpha j}' = iH_{\alpha j}'/(E_\alpha - E_j); \qquad \ldots\ldots(26)$$

the other elements of S' are arbitrary.

The (αj) elements remaining in (24) are of order g^2 at least. Since the Hamiltonian is still Hermitian, term by term in powers of g, we can choose a second canonical transformation $\Psi'(t) = \exp(iS'').\Psi''(t)$ to remove the second-order coupling, where S'' contains the factor g^2, and so on. After n such transformations, the (αj) elements in the Hamiltonian are of order g^{n+1} at least. Systematically ignoring powers of g higher than g^n, we can write

$$i\partial\Psi^n(t)/\partial t = (H^0 + V)\Psi^n(t), \qquad \ldots\ldots(27)$$

where V is Hermitian and has (αβ) and (jk) elements only. The two-nucleon subspace may now be considered separately; in this subspace (27) is a two-particle Schrödinger equation, with the two-nucleon part of V as the potential.

The basic states we refer to in the last sentence, when mentioning the two-nucleon subspace, are not those with which we began. The successive canonical transformations have generated a new set of orthonormal basic states, in 1–1 correspondence with the original set. In terms of the original representation, the new 'two-nucleon' states describe the presence not only of two nucleons, but of mesons and pairs as well. In a certain sense, we have shown how to disentangle the two-nucleon states by recognizing that physical nucleons have clouds of virtual mesons and pairs associated with them, and by including such a cloud in the description of each nucleon. Thus the original representation refers to 'bare' nucleons, and the new one to 'dressed' nucleons. If the electromagnetic field is also considered, we decouple states of two nucleons plus photons from other states; the electromagnetic interactions of 'bare' and 'dressed' nucleons differ.

Generating dressed two-nucleon states is akin to diagonalizing the Hamiltonian in perturbation theory, starting with eigenstates of H^0 and adding corrections in rising powers of g. Here only a partial diagonalization is made.

The unspecified (αβ) and (jk) elements of each S^m bring some arbitrariness into the method. It is reasonable to give them the same symmetry properties as the basic field theory, but otherwise their choice is free. However, the particular choice at any stage affects the definition of the (non-arbitrary) $S_{\alpha j}{}^n$ at later stages, and an unwise choice may give divergent terms in V.†

Each term in V can be represented graphically; take, for example, the second-order term

$$-\tfrac{1}{2}H_{\alpha j}{}'H_{j\beta}{}'(E_\alpha - 2E_j + E_\beta)/(E_\alpha - E_j)(E_\beta - E_j). \qquad \ldots\ldots(28)$$

Suppose that H' is linear in the meson operators, and therefore has matrix elements for creating or annihilating only one meson at a time. Then, with a conventional time-ordering from right to left, we can think of (28) as describing a transition from the two-nucleon state $|\beta)$ to the two-nucleon state $|\alpha)$ via the intermediate state $|j)$ of two nucleons and one meson. According to whether this meson is emitted and absorbed by the same nucleon or by different nucleons, we can represent the process by a diagram like 3(*a*) or 3(*b*) (full directed lines stand for nucleons, broken lines for mesons; later times are above). Virtual pairs, which may appear in higher-order processes, are represented as in figures 3(*d*), 3(*e*). If the arbitrary

† For example, $V_{\alpha\beta}$ contains the term $-\tfrac{1}{2}\Sigma_\gamma S_{\alpha\gamma}{}' S_{\gamma\beta}{}'(E_\alpha - 2E_\gamma + E_\beta)$; the sum (or integral) over γ may diverge.

elements $S_{\alpha\beta}{}^m$ and $S_{jk}{}^m$ all vanish, this representation is sufficient; if not, it will need to be extended.

In the present case, diagrams are only a convenient visualization. In some other methods, however, where similar representations can be set up, there are simple rules for writing down the contribution to the potential from any process, and diagrams become useful aids to computation. The processes which appear, and their contributions to V, depend on the method being used. However, when the adiabatic limit is taken, and the potential is expanded in powers of g (independent of E), there is a real sense in which contributions from each process can be compared in different methods; they prove to be the same.

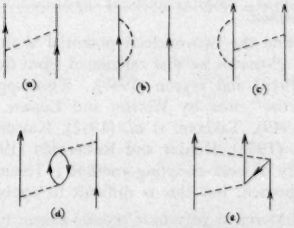


Figure 3. Typical time-ordered diagrams.

In current field theories, certain processes (e.g. figure 3(*b*), (*c*), (*d*)) give divergent contributions unless a cut-off is introduced. A major disadvantage of the present method is that it contains no simple general prescription for removing these divergences, even for a 'renormalizable' theory (in the $\mathscr{S}$-matrix sense).

A certain difficulty has been ignored so far. If the two nucleons are fixed, the denominators $(E_\alpha - E_j)$ in (26) and in the definitions of the other $S_{\alpha j}{}^m$ never vanish. In general, however, we cannot ignore states $|\alpha)$ above the threshold for meson production, and the S^m are not well defined.

In practice, we suppose that some cut-off, which may not be explicitly introduced, prevents virtual nucleon states from occurring with energies much higher than that of the real nucleon states under consideration. The states above meson threshold energy can then be safely ignored when calculating a potential for use at low energies.

There is another way of choosing canonical transformations, best described in the interaction representation (denoted by clarendon type), where (22) becomes

$$i d\mathbf{\Psi}(t)/dt = \mathbf{H}'(t)\,\mathbf{\Psi}(t). \qquad \ldots\ldots(29)$$

The two-nucleon states can be decoupled to first order by the transformation

$$\mathbf{\Psi}(t) = \exp i\mathbf{S}'(t)\,\mathbf{\Psi}'(t), \qquad \ldots\ldots(30)$$

$$\mathbf{S}'(t) = -\int_{-\infty}^{t} \mathbf{H}'(\tau)\,d\tau. \qquad \ldots\ldots(31)$$

The interaction Hamiltonian becomes

$$(\exp -iS')H'(\exp iS') - i(\exp -iS')\frac{\partial(\exp iS')}{\partial t}$$
$$= \tfrac{1}{2}i[H', S'] - \tfrac{1}{3}[[H', S']S'], \ldots, \text{etc.} \qquad \ldots\ldots(32)$$

A sequence of such canonical transformations can eliminate the unwanted coupling terms to any order in g. The difficulty of vanishing denominators does not arise here, but that of renormalization remains. This method can also be framed in the covariant formalism of Tomonaga (1946) and Schwinger (1948) and is so given by Nambu (1948, 1950) and Nishijima (1951).

3.2.2. *The $\mathscr{S}$-matrix method.*

Another way to define the two-nucleon potential is to require it to give the same nucleon–nucleon $\mathscr{S}$-matrix as that calculated from field theory by the techniques of Feynman (1949) and Dyson (1949). This approach is proposed by Nambu (1950) and Bethe (cited by Watson and Lepore 1949); it is used by Watson and Lepore (1949), Taketani *et al.* (1952), Nakabayasi and Sato (1952), Lepore (1952), I. Sato (1953), Henley and Ruderman (1953 b) and Hiida *et al.* (1956). It is essentially a weak-coupling method; Imamura (1955) suggests a somewhat different approach, but this is difficult to apply without returning to perturbation theory.

Consider the scattering of two nucleons by a potential V in the two-body Schrödinger equation. In the interaction representation, the $\mathscr{S}$-matrix has the form

$$\mathscr{S}_{\alpha\beta} = \delta_{\alpha\beta} - 2\pi i\delta(E_\alpha - E_\beta)R_{\alpha\beta} \qquad \ldots\ldots(33)$$

where the matrix R is related to V by

$$R_{\alpha\beta} = V_{\alpha\beta} + V_{\alpha\gamma}(E_\beta - E_\gamma + i\epsilon)^{-1}R_{\gamma\beta}, \qquad \ldots\ldots(34)$$

(Lippmann and Schwinger 1950, Gell-Mann and Goldberger 1953). The matrix notation of § 3.2.1 is used, and an integration over the intermediate state $|\gamma)$ is understood. ϵ is a small, real and positive parameter, made to vanish after the integration is performed; this prescription is to satisfy the boundary conditions for scattering.

When V is known, (34) can be solved to give R and hence $\mathscr{S}$. We would like to invert this process. However, $\mathscr{S}$ defines only the elements of R 'on the energy shell' (for which $E_\alpha = E_\beta$) and does not fully determine V.

Let us solve (34) by iteration for V in terms of R

$$V_{\alpha\beta} = R_{\alpha\beta} - R_{\alpha\gamma}(E_\beta - E_\gamma + i\epsilon)^{-1}R_{\gamma\beta} + \ldots. \qquad \ldots\ldots(35)$$

Writing R and V as power series in g^2, $\Sigma R_{2m}g^{2m}$ and $\Sigma V_{2m}g^{2m}$, we can find the coefficients V_{2m} in succession. In the present case, the matrix elements of R_{2m} off the energy shell are unspecified, and we choose them arbitrarily; thus the corresponding elements of V_{2m} are arbitrary. However, the choice of these elements affects the definition of V_{2n} *on the energy shell* at later stages ($n > m$).

As in the canonical transformation method it is reasonable to give these arbitrary elements the same symmetries as the basic field theory, but otherwise

their choice is free. Again, an unwise choice may give divergent terms in V. Particular prescriptions can be given for defining R off the energy shell (e.g. Henley and Ruderman 1953 b), but this is only a matter of convenience.

The present method gives a whole class of potentials with the same $\mathscr{S}$-matrix. There is, however, no general connection between $\mathscr{S}$ and the bound state energies, except for potentials which decay sufficiently fast at infinity; in this connection see Bargmann (1949 b), Jost and Kohn (1953) and § 2.7.

A great merit of the method is that divergent terms in the $\mathscr{S}$-matrix can be removed by standard techniques, at least in renormalizable theories.

A rather different approach to the $\mathscr{S}$-matrix is also useful. If the matrix elements for the scattering of n mesons by a nucleon are known, both on and off the energy shell, we can deduce the nucleon–nucleon scattering due to the exchange

Figure 4.

of $2n$ virtual mesons (if meson–meson scattering is ignored). The exchange of an odd number of mesons can be deduced from processes in which production or absorption of mesons also occurs.

A simple example is one-meson scattering. In terms of Feynman diagrams, the process of figure 4(b) is inferred from 4(a); the shaded areas represent all meson scattering parts, and 4(b) thus includes the crossing of meson lines. Henley and Ruderman (1953 b) use this approach, estimating meson–nucleon scattering theoretically. Miyazawa (1956) and Klein and McCormick (1956) show more generally that the required meson–nucleon scattering matrix elements can be deduced from experiment, using dispersion relations, if the nucleon is treated as a fixed source with pseudo-vector coupling to the meson field and meson–meson interactions are neglected. The contribution to V from two-meson exchange processes, including radiative corrections, can thus be estimated; this is done explicitly by Konuma *et al.* (1958; § 3.3.6). Novozhilov (1957) and Sharp (1958) use a closely related approach, deriving the processes 4(a) from the Low equation (Low 1955, Wick 1955).

3.2.3. *The Tamm–Dancoff method.*

This method is to find an equation for the two-nucleon part of the state vector by eliminating algebraically all other amplitudes. It is suggested by Tamm (1945) and Dancoff (1950), and is further developed by Lévy (1952 a, b); Feldman (1955) and Klein (1954 a) contribute critical discussions. Królikowski and Rzewuski (1956) use a similar approach.

Consider the time-independent Schrödinger equation

$$H\Psi = E\Psi, \qquad \ldots\ldots(36)$$

where Ψ is an eigenstate of H with eigenvalue E. Let us write Ψ and H symbolically in the block form

$$\Psi = \begin{pmatrix} \psi_0 \\ \psi_1 \end{pmatrix}, \quad H = \begin{pmatrix} H_0, & H_{01} \\ H_{10}, & H_1 \end{pmatrix}. \qquad \ldots\ldots(37)$$

ψ_0 and ψ_1 are the projections of Ψ on the two-nucleon subspace and the complementary subspace, respectively; H_0 is the two-nucleon part of H, and so on. We now have two coupled equations,

$$\left.\begin{aligned} (E-H_0)\psi_0 &= H_{01}\psi_1 \\ (E-H_1)\psi_1 &= H_{10}\psi_0 \end{aligned}\right\}. \qquad \ldots\ldots(38)$$

Formally eliminating ψ_1, we find

$$(E-H_0)\psi_0 = H_{01}(E-H_1)^{-1}H_{10}\psi_0, \qquad \ldots\ldots(39)$$

where $(E-H_1)^{-1}$ denotes the appropriate Green's function. This looks like a two-nucleon Schrödinger equation, with the right-hand side as the potential (assuming there are no interaction terms in H_0). The Tamm–Dancoff method is to identify it as the equation of physical interest.

This method may be used with the assumption, known as the *Tamm–Dancoff approximation*, that all amplitudes in ψ_1 describing more than a certain number of mesons and pairs can be ignored. Equation (38) then reduces to a finite number of coupled integral equations which one may hope to solve exactly. This appears to offer an improvement over perturbation theory, but meets difficulties in practice (Klein 1954 a).

On the other hand, (39) may be used as the start of a weak-coupling approximation. Let us write $H_1 = H_1^0 + H_1'$, in terms of the free-field and interaction parts, and use the formal identity

$$(a-b)^{-1} = a^{-1} + a^{-1}ba^{-1} + a^{-1}ba^{-1}ba^{-1} + \ldots \qquad \ldots\ldots(40)$$

to expand the Green's function $(E-H_1)^{-1}$:

$$(E-H_1^0-H_1')^{-1} = (E-H_1^0)^{-1} + (E-H_1^0)^{-1}H_1'(E-H_1^0) + \ldots. \qquad \ldots\ldots(41)$$

This expression has no singularities when E is below the meson production threshold. Introducing the projection operator on to the two-nucleon subspace P, and the complementary projection operator $Q = 1-P$, the 'potential' in (39) may now be written

$$PH'\left\{\frac{Q}{E-H^0} + \frac{Q}{E-H^0}H'\frac{Q}{E-H^0} + \frac{Q}{E-H^0}H'\frac{Q}{E-H^0}H'\frac{Q}{E-H^0} + \ldots\right\}H'P \qquad (42)$$

$$= PH'\frac{Q}{E-H^0}\left(1-H'\frac{Q}{E-H^0}\right)^{-1}H'P. \qquad \ldots\ldots(43)$$

This expression is given by Fukuda *et al.* (1954), among others. It is convenient to describe the terms in (42) by time-ordered diagrams, as in § 3.2.1 (Lévy 1952 a, b); there are simple rules for writing down the contributions corresponding to given diagrams and the latter are very helpful in computations.

Now there are certain objections to identifying (39) as the physical two-nucleon Schrödinger equation. In a conventional Schrödinger equation the potential is Hermitian (to conserve probability) and independent of the eigenvalue E (E-dependent potentials do not have orthogonal eigenstates in general). However, the 'potential' here is explicitly E-dependent. We can eliminate E from the right-hand side of (39) by iteration, but the resulting modified potential—henceforth denoted by V_{TD}—is no longer Hermitian (Feldman 1955, Fukuda *et al.* 1954). Lévy (1952 a, b) neglects nucleon motion, and in this approximation the non-Hermitian terms happen to vanish, but in general the objection remains. This does not imply that the equations are somehow wrong; it simply means that we have not yet found anything that can be identified as a conventional two-nucleon Schrödinger equation.

We would like to pass from the present equations to some of a more familiar form. It is instructive to make a comparison with the canonical transformation method; in the latter the eigenstate of H^0+H' describing two physical ('dressed') nucleons is *transformed* into the two-nucleon subspace, but in the Tamm–Dancoff method it is *projected* into this subspace.

Suppose that the two-nucleon subspace has been decoupled by a canonical transformation U, and choose a complete orthonormal set of states $\psi(A)$ in it which satisfy the resulting two-nucleon Schrödinger equation

$$(H^0+V)\psi(A) = E(A)\psi(A), \qquad \ldots\ldots(44)$$

where A is some convenient label. Now the eigenstates $\Psi(A) = U^{-1}\psi(A)$ of the Hamiltonian H^0+H' are not confined to this subspace; their two-nucleon projections—denoted by $\psi_0(A)$—satisfy the Tamm–Dancoff equation

$$(H^0+V_{\text{TD}})\psi_0(A) = E(A)\psi_0(A). \qquad \ldots\ldots(45)$$

If the $\psi_0(A)$ are a complete set in the two-nucleon subspace, which is certainly true in the weak-coupling limit, they are related to the $\psi(A)$ by a non-singular linear operator N, $\psi_0(A) = N\psi(A)$; N is not unitary in general. It follows that

$$N^{-1}(H^0+V_{\text{TD}})N = H^0+V. \qquad \ldots\ldots(46)$$

The Tamm–Dancoff Hamiltonian can thus be made Hermitian by a suitable (non-canonical) transformation. N is easily found to low orders by perturbation theory.

Fukuda *et al.* (1954) and Okubo (1954) study this transformation in more detail. Taking a complete orthonormal set of 'two-dressed-nucleon' states $\Psi(A)$ (which we may suppose to be defined as above), and making the same assumptions as before, they introduce the linear operator J operating between the two-nucleon subspace and the full space, defined by the property $\Psi(A) = J\psi_0(A)$. If J^* is the Hermitian conjugate of J, J^*J is a linear operator in the two-nucleon subspace. Assuming that a Hermitian operator $(J^*J)^{1/2}$ exists and has an inverse, it follows that the states $\chi_0(A) = (J^*J)^{1/2}\psi_0(A)$ are a complete orthonormal set in the two-nucleon subspace. It is proposed to use the $\chi_0(A)$ as physical two-nucleon wave functions, and the appropriate transform of (45) as the corresponding Schrödinger equation. The resulting two-nucleon Hamiltonian is

$$H^0+V = (J^*J)^{-1/2}[J^*(H^0+H')J](J^*J)^{-1/2} \qquad \ldots\ldots(47)$$

which proves to be the same as that found by canonical transformations. When the nucleons are treated as fixed sources, this is simply the potential ; the perturbation expansions in this limit are

$$J^*(H^0+H')J = -PH'\frac{Q}{H^0}H'P-PH'\frac{Q}{H^0}H'\frac{Q}{H^0}H'\frac{Q}{H^0}H'P-\ldots, \quad \ldots\ldots(48)$$

$$(J^*J)^{\frac{1}{2}} = 1+\tfrac{1}{2}PH'\frac{Q}{(H^0)^2}H'P+\ldots, \quad \ldots\ldots(49)$$

assuming H' to be linear in the meson operators, and normalizing H^0 to be zero in the state of two bare nucleons ($PH^0 = 0$). $J^*(H^0+H')J$ is called the 'normal part' and the factors $(J^*J)^{-\frac{1}{2}}$ the 'probability part'; in the fixed-source limit they commute. $(J^*J)^{-1}$ is simply the probability that no mesons are present ; it should lie between 0 and 1, which is a useful guide to the region of validity of the expansions. Fukuda *et al.* (1954) suggest that these parts should be treated separately ; a potential based on separate expansions is calculated by Inoue *et al.* (1956 ; § 3.3.6). S. Sato (1955) and I. Sato *et al.* (1955) make further calculations.

Unfortunately, the Tamm–Dancoff method cannot be renormalized. Hoping to avoid this difficulty, Cini (1953) proposes a covariant reformation of the method, and Dyson (1953) suggests a variant in which the two-nucleon amplitudes are defined differently (the 'new Tamm–Dancoff' method). Neither method can be completely renormalized (Lehmann 1953, Taylor 1954 a), although there is an approximation to the 'new' method which can (Taylor 1954 a).

The new Tamm–Dancoff method is further discussed by Klein (1954 d), and applied to nuclear forces by Taylor (1954 b) and Kurşunoğlu (1954). In the adiabatic limit (§ 3.3.1) which is usually taken, the potential is the same as that of the original Tamm–Dancoff method.

3.2.4. *The Lévy–Klein method.*

This method approaches nuclear forces through the relativistic two-body equation (Salpeter and Bethe 1951, Schwinger 1951, Gell-Mann and Low 1951) by reducing the latter to the form of a Schrödinger equation. Lévy (1952 a) shows that in many cases this gives the same result as the Tamm–Dancoff method. Since the Bethe–Salpeter equation is renormalizable, one can thus treat divergent processes for which the Tamm–Dancoff methods fails ; Lévy (1952 b) uses both methods together in his discussion of nuclear forces.

The reduction of the Bethe–Salpeter equation is examined more rigorously by Klein (1953 b, 1954 b) and Macke (1953 a, b, c). They show the correspondence with the Tamm–Dancoff results to be close but not exact : for example, vacuum processes do not appear, and there are differences in certain energy denominators. The subject is also discussed by Symanzik (1954) and Zimmerman (1954). There is recent work on this kind of approach by Klein and McCormick (1958).

The Bethe–Salpeter equation for two nucleons can be written

$$(i\tfrac{1}{2}K_\mu\gamma_\mu{}^a+ip_\mu\gamma_\mu{}^a+M)(i\tfrac{1}{2}K_\mu\gamma_\mu{}^b-ip_\mu\gamma_\mu{}^b+M)\psi(p_\mu)$$
$$= \int d^4k G(k_\mu)\psi(p_\mu+k_\mu), \quad \ldots\ldots(50)$$

where the superfixes a, b distinguish the nucleons. p_μ and K_μ are the relative and total 4-momenta, respectively; the 'centre-of-mass' frame $K_\mu = (0, 0, 0, iE)$ is generally used. The γ-matrices are defined in terms of the Dirac matrices by $\gamma_r = -i\beta\alpha_r$ $(r = 1, 2, 3), \gamma_4 = \beta$. The kernel $G(k_\mu)$ is a power series in g, the individual terms of which are finite (renormalized) and can be represented by Feynman diagrams. The wave function $\psi(p_\mu)$ is a 16-component spinor.

In coordinate space, ψ depends on the relative space and time $x, x_4 = ict$. For a three-dimensional description, we need a wave function describing two particles at the *same* time, and we therefore try to find an equation for $\psi(x, 0) \equiv \phi(x)$.

Now in momentum space

$$\phi(\mathbf{p}) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \psi(p_\mu)\, dp_4. \qquad \ldots\ldots(51)$$

To reduce (50) to an equation for $\phi(\mathbf{p})$, we must know how $\psi(p_\mu)$ depends on p_4, and this dependence must be separable.

Salpeter (1952) and Lévy (1952 a) show that when only the g^2 term is retained in G ('ladder approximation': H' assumed linear in meson field operators) and retardation effects are ignored, (50) reduces at once to a three-dimensional equation. The dependence of $\psi(p_\mu)$ on p_4 is known in this case. We now assume the same dependence in the general case, as a zeroth approximation; higher approximations are to be found by iterating the equation for $\psi(p_\mu)$. Klein (1954 b) stresses that in general there exists no single-time equation for the exact covariant wave function ψ: we have to assume that its dependence on p_4 can be approximately separated in this way.†

Now ϕ has sixteen spinor components, whereas a wave function for two non-relativistic nucleons should have only four. Consider a representation in terms of free-particle solutions of the Dirac equation, which have both positive and negative energies of course. We may divide ψ into four parts $\psi_{++}, \psi_{+-}, \psi_{-+}, \psi_{--}$, which are its projections on to the subspaces where both particles have positive energies (ψ_{++}), particle a has positive energy and particle b has negative energy (ψ_{+-}), etc. Each part has only four spinor components.

Equations (50), (51) and the assumed p_4-dependence of ψ are now used to derive an equation for ϕ_{++} of the form

$$(H^0 + V_{LK})\phi_{++} = E\phi_{++}, \qquad \ldots\ldots(52)$$

eliminating the unwanted components ϕ_{+-}, etc. by perturbation theory. The 'potential' V_{LK} is accurate to a given order in g.

However, V_{LK} contains E explicitly; if this E-dependence is removed by iterating (52), the modified potential is not Hermitian. In fact there are the same objections here as in the Tamm–Dancoff case. Except in certain approximations where the non-Hermitian terms vanish, such as the adiabatic limit used by Lévy (1952 b) and Klein (1953 b), this is not a conventional Schrödinger equation.

Arnowitt and Gasiorowicz (1954) question the elimination of the unwanted components ϕ_{+-}, etc. by perturbation theory; their results suggest this approach may be misleading at small internucleon distances. Green and Biswas (1957), working to order g^2 only, give another treatment which avoids this feature.

† However, see Królikowski and Rzewuski (1955).

3.2.5. *The Brueckner–Watson method.*

The method of Brueckner and Watson (1953 a, b) is to set up the Møller scattering matrix for the full field theory in the formalism of Lippmann and Schwinger (1950), and to find the submatrix for two-nucleon scattering by algebraic elimination; certain terms are then identified as the two-nucleon potential. This procedure is rather lengthy to explain, however. We give an alternative derivation, suggested by Klein (1953 c, 1954 a), which depends on an equivalent algebraic elimination.

Assume for convenience that H' is linear in the meson operators, and that the creation of nucleon–antinucleon pairs can be ignored (the generalization will be evident). Then the Hamiltonian eigenstates in which we are interested only have components describing two nucleons plus n mesons ($n \geqslant 0$), and we can divide each eigenstate Ψ into an 'odd' part Ψ_o and an 'even' part Ψ_e corresponding to n odd and even. Hence the time-independent Schrödinger equation can be decomposed

$$\left.\begin{aligned}(E-H^0)\Psi_e &= H'\Psi_o\\ (E-H^0)\Psi_o &= H'\Psi_e\end{aligned}\right\} \quad \ldots\ldots(53)$$

Eliminating Ψ_o, $\quad (E-H^0)\Psi_e = H'(E-H^0)^{-1}H'\Psi_e \equiv \Delta_0\Psi_e$. $\quad \ldots\ldots(54)$

$(E-H^0)^{-1}$ denotes the appropriate Green's function; it has no singularity when E is below meson production threshold. We shall not discuss the singularities of the more complicated Green's functions which appear later.

Now Ψ_e has components describing two nucleons plus $2n$ mesons, and we can divide it into odd and even parts Ψ_{o1} and Ψ_{e1} corresponding to n odd and even. Since H' creates or annihilates one meson, we can separate

$$\Delta_0 \equiv H'(E-H^0)^{-1}H'$$

into a 'non-diagonal' part u_0, which creates or annihilates two mesons, and a 'diagonal' part v_0 which does not change the number of particles. The equation (54) can now be decomposed in analogy with (53), and Ψ_{o1} eliminated to give

$$(E-H^0-v_0)\Psi_{e1} = u_0(E-H^0-v_0)^{-1}u_0\Psi_{e1} \equiv \Delta_1\Psi_{e1}. \quad \ldots\ldots(55)$$

It is clear how the process continues. At each stage we divide the interaction term into diagonal and non-diagonal parts, and the state function into odd and even parts; this odd part is then eliminated. Repeating the cycle m times, we have the equation

$$(E-H^0-v_0-v_1\ldots v_{m-1})\Psi_{em} = \Delta_m\Psi_{em}, \quad \ldots\ldots(56)$$

in an obvious notation. If the process converges as $m\to\infty$, Δ_m becomes vanishingly small, Σv_i from 0 to ∞ converges, and Ψ_{em} becomes the two-nucleon projection of Ψ, with zero elements elsewhere. Writing V_{BW} for the two-nucleon submatrix of Σv_i from 0 to ∞, and ψ_0 for the two-nucleon projection of Ψ, the limit of (56) is

$$(E-H^0-V_{BW})\psi_0 = 0, \quad \ldots\ldots(57)$$

which looks like a two-nucleon Schrödinger equation.

However, V_{BW} depends explicitly on E. If this E-dependence is eliminated by iterating (57), the potential becomes non-Hermitian, as in the Tamm–Dancoff and Lévy–Klein methods; see the discussion in § 3.2.3.

Indeed, there is a close correspondence between the Tamm–Dancoff and Brueckner–Watson methods. Both derive equations for the two-nucleon projections of the eigenstates Ψ, though the unwanted components are not eliminated in the same order. If all E-dependence is removed from the potential in each case, the results must be identical, although the terms do not appear in the same succession.

The lowest approximation to V_{BW} is the two-nucleon submatrix of v_0, which is precisely the lowest Tamm–Dancoff term (compare (42)). It accounts for processes in which one meson is emitted and later absorbed by the factors H', while in the intermediate state the system propagates freely according to the Green's function $(E-H^0)^{-1}$. The second term in V_{BW} is the two-nucleon submatrix of v_1 (the diagonal part of $u_0(E-H^0-v_0)^{-1}u_0$). It accounts for processes in which two nucleons are emitted and later absorbed by the factors u_0, while in the intermediate state the system of two nucleons and two mesons propagates according to the Green's function $(E-H^0-v_0)^{-1}$. Now v_0 has matrix elements for the scattering of two nucleons and two mesons; this Green's function therefore includes the effects of multiple scattering by v_0 in the intermediate state. v_1 contains a selection of terms of arbitrarily high order in the coupling constant; the series Σv_i is not a simple perturbation series.

Tani (1954) suggests that the odd parts should be eliminated at each stage by canonical transformations instead, to ensure a Hermitian potential.

We add finally that the Brueckner–Watson method contains no simple general prescription for treating divergent terms.

3.2.6. *Uniqueness.*

When nucleon motion is neglected, we expect the potential found by any method to reduce to the potential energy of fixed sources (the 'adiabatic approximation'; § 3.1). When nucleon motion is taken into account, however, we have seen in § 2.2 that many potentials can give the same binding energies and scattering. It is therefore not surprising that in this case the potentials found from meson field theory are not unique.

In the $\mathscr{S}$-matrix and canonical transformation methods, the arbitrariness enters quite openly; certain elements are unspecified and can be chosen freely. The other methods seem at first to allow no arbitrariness. However, we are always free to make a canonical transformation of the final two-nucleon equation—or, for that matter, of the original nucleon–meson field theory—and either procedure gives fresh potentials. The particular choice of potential is ultimately a matter of convenience.

In the canonical transformation method, the choice in which all the arbitrary matrix elements $S_{\alpha\beta}{}^m, S_{jk}{}^m$ vanish appears the most natural. In this case the modifications to non-mesonic interactions, brought about by the decoupling transformations, can all be represented in terms of virtual meson processes.

An alternative criterion of convenience is that the potential should be as nearly static as possible. Another requirement might be that the three-nucleon (or more

generally, many-nucleon) potential, which can be constructed in a similar way and is itself somewhat arbitrary, shall be as nearly as possible a sum of two-nucleon potentials. In practice, however, the chosen criterion is usually mathematical simplicity.

3.3. *Potentials from the Pion Field*

3.3.1. *The pion field.*

Many properties of the pions (or π-mesons) found in nature are now well known (Gell-Mann and Watson 1954, Bethe and de Hoffmann 1955). π^+, π^- and π^0 are pseudoscalar particles with positive, negative and zero unit charges respectively. They interact strongly with nucleons; their other interactions are usually ignored in the nuclear force problem. The pion masses are $\mu^{\pm} \simeq 273\, m_e$, $\mu^0 \simeq 264\, m_e$; the differences are presumed to have an electromagnetic origin (Feynman and Speisman 1954) and in practice pions are generally assumed to have the same mass $\mu \simeq 270\, m_e$, $\mu^{-1} \simeq 1{\cdot}4 \times 10^{-13}$ cm. The possible existence of a second neutral pion (Baldin 1958, Yamaguchi 1958) is ignored throughout this section.

We represent the charged pions by a complex pseudoscalar field

$$\phi = (1/\sqrt{2})(\phi_1 + i\phi_2),$$

with real components $\phi_1(x)$ and $\phi_2(x)$, and the neutral pion by a real pseudoscalar field $\phi_3(x)$. The coupling to the nucleon field $\psi(x)$ is usually supposed to be symmetrical between the three fields ϕ_j, in the manner first suggested by Kemmer (1938); such couplings give charge-independent nuclear forces. There are many possible symmetrical couplings, but only two are generally considered. These have the interaction Lagrangian densities

$$-ig \Sigma_j \bar{\psi}(x) \gamma_5 \tau_j \psi(x) \phi_j(x), \qquad \ldots\ldots(58)$$

$$-if\mu^{-1} \Sigma_{j,\nu} \bar{\psi}(x) \gamma_5 \gamma_\nu \tau_j \psi(x)\, \partial\phi_j(x)/\partial x_\nu, \qquad \ldots\ldots(59)$$

where $\bar{\psi} = \psi^* \gamma_4$, the γ_ν are defined as in § 3.2.4, $\gamma_5 = \gamma_1\gamma_2\gamma_3\gamma_4$, and the τ_j are the usual isobaric-spin matrices. Expression (58) is the direct or pseudoscalar coupling; (59) is the gradient or pseudovector coupling. g and f are dimensionless coupling constants.

For a pseudoscalar meson theory to be renormalizable in the standard perturbation-theoretical sense, the coupling must be pseudoscalar (Matthews and Salam 1951), which is a strong practical argument for this form. However, though pseudovector coupling is not fully renormalizable in this sense,† it can account for much data as well as the other kind.

The pseudovector interaction Hamiltonian density is

$$\frac{f}{\mu} \psi^* [\boldsymbol{\sigma} . \nabla(\boldsymbol{\tau} . \boldsymbol{\varphi}) - \gamma_5 \boldsymbol{\tau} . \boldsymbol{\pi}] \psi + \frac{f^2}{2\mu^2} (\psi^* \gamma_5 \boldsymbol{\tau} \psi)^2, \qquad \ldots\ldots(60)$$

in an obvious vector notation, where $\pi_j(x)$ is canonically conjugate to $\phi_j(x)$, and $\sigma_r = i\gamma_4\gamma_5\gamma_r$, $(r = 1, 2, 3)$. The first two terms in (60) describe pion–nucleon coupling; the third is a nucleon contact interaction.

† For a sense in which it may be renormalizable, see Arnowitt and Deser (1955), Cooper (1955) and Green (1956).

It is often useful to consider the limit where nucleons are regarded as fixed sources of the pion field. (60) is not easy to interpret directly in this limit, for the matrix γ_5 is 'odd' (i.e. it couples positive- and negative-energy nucleon states) and favours transitions in which a nucleon–antinucleon pair takes part. However, Foldy and Wouthuysen (1950) show that odd components can be eliminated from the Hamiltonian to any order in $1/M$ (M = nucleon rest mass) by canonical transformations. When this is done, (60) becomes

$$\frac{f}{\mu}\psi^*\left\{\boldsymbol{\sigma}.\boldsymbol{\nabla}(\boldsymbol{\tau}.\boldsymbol{\varphi})+\gamma_4\frac{[\boldsymbol{\sigma}.\mathbf{p},\boldsymbol{\tau}.\boldsymbol{\pi}]_+}{2M}+\gamma_4\frac{f}{\mu}\frac{\boldsymbol{\pi}^2}{2M}\right\}\psi \qquad \ldots\ldots(61)$$

to order $1/M$, omitting nucleon contact terms.† $\mathbf{p}$ is the nucleon momentum, operating on ψ^* and ψ. The last term is a pair effect, coming from the elimination of virtual pair transitions. We now define the 'adiabatic' limit by restricting ourselves to positive-energy nucleon states and letting M tend to ∞. In this limit (61) reduces to the first term, and the two-nucleon potential becomes the potential energy between fixed sources (the 'adiabatic approximation' of § 3.1) coupled to the pion field by $(f/\mu)\boldsymbol{\sigma}.\boldsymbol{\nabla}(\boldsymbol{\tau}.\boldsymbol{\varphi})$. $\boldsymbol{\sigma}$ is now the ordinary 2×2 spin operator.

The pseudoscalar interaction Hamiltonian density is simply (58) with a change of sign. The odd matrix γ_5 appears here too. If odd components are again eliminated by canonical transformations (Dyson 1948, Foldy 1951, Tani 1951, Berger *et al.* 1952), it becomes

$$\frac{g}{2M}\psi^*\{\boldsymbol{\sigma}.\boldsymbol{\nabla}(\boldsymbol{\tau}.\boldsymbol{\varphi})+\frac{1}{2M}[\boldsymbol{\sigma}.\mathbf{p},\boldsymbol{\tau}.\boldsymbol{\pi}]_+ +\gamma_4 g\boldsymbol{\varphi}^2+\frac{g}{2M}\boldsymbol{\tau}.\boldsymbol{\varphi}\times\boldsymbol{\pi}\}\psi, \qquad \ldots\ldots(62)$$

to order $1/M^2$, omitting nucleon contact terms. The first two terms are simply the pseudovector interaction (61) to this order, with $f/\mu = g/2M$. The last two are 'pair' terms, coming from the elimination of virtual pair transitions. There is evidence that some pair effects are damped by higher order processes, and that at least the $\boldsymbol{\varphi}^2$ term is largely suppressed (§ 3.3.8). Pseudoscalar and pseudovector coupling thus become very similar.

We must define an adiabatic limit with care for pseudoscalar coupling. If we simply let $M\to\infty$, (62) vanishes. If g^2/M tends to a finite limit as $M\to\infty$, the $\boldsymbol{\varphi}^2$ term alone remains (there are no higher terms of order $(g^2/M)^n$). This is not a very useful limit, however, for pair processes do not in fact appear to dominate the pion–nucleon interaction. A more useful limit can be defined as follows, *for the untransformed γ_5-coupling*. In calculating the potential, nucleon and antinucleon kinetic energies are ignored compared to M; the matrix element of (58) between a nucleon state of momentum $\mathbf{p}_1$ and a state of momentum $\mathbf{p}_2$ (or for annihilation with an anti-nucleon of momentum $-\mathbf{p}_2$) is written in terms of $\mathbf{p}_1+\mathbf{p}_2$ and $\mathbf{p}_1-\mathbf{p}_2$; $\mathbf{p}_1+\mathbf{p}_2$ is put equal to zero and $\mathbf{p}_1-\mathbf{p}_2$ is ultimately integrated, being the momentum of the coupled pion. We take this prescription to define the adiabatic limit for pseudoscalar coupling; several 'adiabatic' limits are distinguished in the literature, but this is essentially the one used whenever

† They contribute to the potential terms of zero range, in which we shall not be interested: see § 3.3.2.

potentials are calculated. The two-nucleon potential in this limit is static and unique.†

This adiabatic potential is usually calculated to particular orders in g^2 and μ/M.‡ If only terms of order $g^2\mu^2/M^2$, $g^4\mu^2/M^2$ and $g^4\mu^3/M^3$ are considered, it is the potential energy of fixed sources coupled to the pion field by

$$[g/2M\boldsymbol{\sigma}.\boldsymbol{\nabla}(\boldsymbol{\tau}.\boldsymbol{\varphi})+g^2/2M\boldsymbol{\varphi}^2],$$

calculated to the same order. The latter is often taken to represent pseudoscalar coupling for fixed nucleons. More terms must be included in it when higher orders in g^2 and μ/M are considered; however, radiative corrections and the question of renormalization now appear, and it is generally easier to abandon the fixed-source model and work with the complete γ_5 theory in the adiabatic limit.

3.3.2. *The form of the potential.*

The methods of § 3.2 lead in general to a non-local potential, which may be written as an integral operator on the relative coordinate $\mathbf{r}$ or momentum $\mathbf{p}$, or simply as a function of both $\mathbf{r}$ and $\mathbf{p}$. Taking the latter form, we shall restrict ourselves for convenience to static terms and those linear in $\mathbf{p}$. Because of the symmetries of the pion field theories we consider, the only possible linear dependence on $\mathbf{p}$ is through an $\mathbf{L}.\mathbf{S}$ spin–orbit term (see the discussion in § 2.2).

At distances much larger than the nucleon Compton wavelength M^{-1}, we expect the interaction to be dominated by the exchange of mesons with low momentum ($\ll M$) and nucleon recoil effects to be small. Here the potential should differ little from that found in the adiabatic limit§, which is static. On the other hand, we may expect strong non-adiabatic effects at distances up to a few times M^{-1}.

Now n-pion exchange processes make a contribution of characteristic range $(n\mu)^{-1}$ (Wick 1938). At large distances the one-pion exchange processes dominate; at shorter distances two-pion exchange also becomes important. At the characteristic range of three-pion exchange effects ($r = 1/3\mu \simeq 2/M$) we expect very large non-adiabatic effects. At shorter distances the exchange of K-mesons (mass about $3{\cdot}5\mu$) may also be significant; conservation of the 'strangeness' quantum number makes two-K exchange the lowest strong process.

Since we lack the mathematical power to calculate the potential exactly, a reasonable approach is to approximate it outside some critical radius $r_c (r_c \sim 1/2\mu)$ by the adiabatic one- and two-pion exchange potentials, and to represent it phenomenologically inside this radius. Higher-order and non-adiabatic effects should then be studied for $r > r_c$, and included if necessary.

Such a potential should be applicable to the deuteron, and to low-energy scattering (relative momenta $p \ll M$; 20 MeV lab. energy corresponds to $p = 0{\cdot}1M$),

† The arbitrary parts vanish. For example, the term shown in a footnote to § 3.2.1 vanishes because $E_\alpha = E_\beta = E_\gamma = 2M$. However, if the original γ_5 coupling is first canonically transformed, it is not obvious that our prescription will give the same potential.

‡ It is better here to speak of orders in $\mu/M \simeq 0{\cdot}15$; M appears only in this dimensionless ratio provided potential depth and any cut-off momenta are expressed in units μ, and distances in units μ^{-1}. The μ/M expansion is found in practice by expanding the integrand of each contribution before integrating over virtual pion momenta. Many authors criticize this procedure, for the expansion is usually valid over only part of the range of integration, but there has been no practical alternative.

§ The non-relativistic approximations generally made are presumably justified here.

for which the details of the inner phenomenological region are not important. We cannot be equally confident at higher energies.

The adiabatic potentials have inadmissible singularities at $r = 0$; the present method avoids this difficulty. The sum of contributions from higher-order processes may not converge everywhere; it is sufficient here if they converge for $r > r_c$.

We have been rather optimistic, above, and in fact not all our expectations are fulfilled (see especially § 3.3.7). However, this approach—pioneered by Taketani *et al.* (1951, 1952) and by Lévy (1952 b)—achieves a fair measure of success.

3.3.3. *The phenomenological inner region.*

The inner region $r \leqslant r_c$ is treated phenomenologically, because of our ignorance of higher-order and non-adiabatic effects at short distances (where weak-coupling theory may diverge anyway). The actual size of such effects (§ 3.3.7) makes us increase our first estimate of r_c to at least $r_c \sim 0{\cdot}7\mu^{-1}$.

The details of this phenomenological treatment should be unimportant for low-energy interactions. A square-well potential with empirically determined radius $r_0 < r_c$, depth and spin- and parity-dependence may be used. In practice the extreme case of a 'hard core' (infinitely repulsive well) is usually taken, partly because of its success in explaining high-energy p–p scattering (§ 2.4.4). The meson-theoretical outer potential is continued into the region $r_0 < r < r_c$, but this is a purely empirical procedure.

Using a non-adiabatic approximation appropriate for small distances, Lévy (1952 b) shows that the δ-function in the lowest order adiabatic potential (§ 3.3.4) is replaced by a strongly repulsive short-range non-local potential, acting only in S-states.† This encourages the belief‡ in a hard core.

Gartenhaus (1955; § 3.3.6) introduces no separate inner potential, but puts an empirical cut-off in the pion–nucleon coupling. His potential is unchanged asymptotically, but depends strongly on the cut-off at small distances. This method proves adequate for S-wave scattering and the deuteron. The cut-off spreads the δ-function mentioned above into a strong short-range local potential, repulsive in even-parity states but attractive in odd-parity states.

Other treatments of the inner region suggest themselves, for example an energy-independent boundary condition at a certain radius (used in other contexts by Breit and Bouricius (1949) and Feshbach and Lomon (1956); § 2.5). It might be convenient to introduce the short-range spin–orbit force suggested by high-energy scattering (§§ 2.4.6, 2.6, 2.8, 3.3.10) by a boundary condition, or as part of an empirical inner potential. Such possibilities have not yet been explored.

3.3.4. *The one-pion exchange potential.*

The leading contribution to this potential, in powers of g, comes from the process shown in figure 5(*a*). In the adiabatic limit, for either coupling ($f/\mu = g/2M$) it is

$$V^{(2)}(r) = \tfrac{1}{3}\mu\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}\frac{f^2}{4\pi}\left\{\left[\boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)} + S_{12}\left(1+\frac{3}{x}+\frac{3}{x^2}\right)\right]\frac{e^{-x}}{x} + \boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}\,\delta(x)\right\}, \quad \ldots\ldots(63)$$

† To be precise, a separable potential: see §§ 2.2, 2.4.7.

‡ Various other theoretical arguments for a repulsive core are advanced by Werle (1955) and Green and Biswas (1957).

where $x = \mu r$, $S_{12} = 3(\boldsymbol{\sigma}^{(1)}.\mathbf{r})(\boldsymbol{\sigma}^{(2)}.\mathbf{r})/r^2 - \boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}$, and $\boldsymbol{\sigma}^{(i)}, \boldsymbol{\tau}^{(i)} (i = 1, 2)$ are the Pauli spin and isobaric-spin operators of the two nucleons.

The δ-function in (63) has no effect at finite distances, and can be omitted. The other terms are a spin-dependent central potential and a tensor potential, both isobaric spin-dependent.

Contributions of higher order in g come from the processes shown in figure 5(b), where the shaded areas represent all radiative corrections; all possible time-orderings are understood. For pseudoscalar coupling Hiida *et al.* (1956) show

Figure 5.

that at large distances all these contributions are included in (63) if the renormalized† coupling constant is used (we always take μ and M to be the physical, i.e. renormalized, masses). Pseudovector coupling cannot be fully renormalized in general, but for fixed-source theory with a cut-off the same result holds (Miyazawa 1956, Klein and McCormick 1956).

The tensor potential has the right sign to explain the deuteron quadrupole moment, which was a reason given for postulating pseudoscalar mesons even before pions were discovered (Pauli 1946). The central potential is repulsive in the triplet odd state, and helps to explain the low-energy ^{3}P-wave anomaly (Otsuki and Tamagaki 1954, 1955); see § 2.4.3.

3.3.5. *The two-pion exchange potential.*

The fixed source models described in § 3.3.1 are used to illustrate the leading terms in the adiabatic limit. To orders $g^4(\mu/M)^2$ and $g^4(\mu/M)^3$, pseudoscalar coupling gives the processes shown in figure 6(a), (b), which correspond to two-pair and one-pair processes in the original γ_5 theory. To order f^4, pseudovector

Figure 6.

coupling gives the processes 6(c), (d). All possible time-orderings of the meson lines are to be understood. We do not press the fixed-source model to higher orders, for the reasons given in § 3.3.1.

Figure 6(c) includes a class of time-ordered diagrams (e.g. figure 7(a)) in which there is an intermediate state of two nucleons alone, and a complementary

† The definition of the renormalized coupling constant is somewhat arbitrary (Deser *et al.* 1954). Unless otherwise stated, we shall mean the constant defined by the coupling of a free nucleon to a pion of zero 4-momentum.

class (e.g. figure 7(*b*)) in which there is not. In the $\mathscr{S}$-matrix and canonical transformation methods, and in Rayleigh–Schrödinger perturbation theory for fixed nucleons, both classes appear directly in fourth order. On the other hand, in the Tamm–Dancoff, Lévy–Klein and Brueckner–Watson methods, and in

Figure 7.

Brillouin–Wigner perturbation theory for fixed nucleons, processes like 7(*a*) do not appear directly in fourth order. However, the latter methods give an E-dependence† in second order ; when this is eliminated by iteration, the missing fourth-order terms appear.

The 'two-pair' processes 6(*a*) contribute, in the adiabatic limit,

$$V_{2p}^{(4)}(r) = -6\mu\left(\frac{g^2}{4\pi}\right)^2\left(\frac{\mu}{2M}\right)^2\frac{\mathrm{K}_1(2x)}{\pi x^2}, \qquad \ldots\ldots(64)$$

where K_n is the nth-order Hankel function of imaginary argument (Watson 1944) and $x = \mu r$ as before. This is just the fourth-order potential of fixed-source scalar meson pair theory (Wentzel 1942), with the appropriate factor for symmetrical pion coupling.

The 'one-pair' processes 6(*b*) contribute

$$V_{1p}^{(4)}(r) = 6\mu\left(\frac{g^2}{4\pi}\right)^2\left(\frac{\mu}{2M}\right)^3(1+x^2)^2x^{-6}\mathrm{e}^{-2x}. \qquad \ldots\ldots(65)$$

The pseudovector processes like 7(*a*) contribute

$$\begin{aligned}V^{(4)\prime}(r) = {} & \left(\frac{f^2}{4\pi}\right)^2(3-2\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)})\frac{2\mu}{\pi x^3}\\ & \times \mathrm{e}^{-x}\{(2+2x+x^2)\,\mathrm{K}_0(x)+(4+4x+x^2)\,x^{-1}\,\mathrm{K}_1(x)\\ & +\tfrac{2}{3}\boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}\,[(1+x)\,\mathrm{K}_0(x)+(2+2x+x^2)\,x^{-1}\,\mathrm{K}_1(x)]\\ & -\tfrac{1}{3}S_{12}[(1+x)\,\mathrm{K}_0(x)+(5+5x+x^2)\,x^{-1}\,\mathrm{K}_1(x)]\},\end{aligned} \qquad \ldots\ldots(66)$$

omitting a δ-function term.

The contribution $V^{(4)\prime\prime}(r)$ from the remaining pseudovector processes 7(*b*) contains terms which cancel against $V^{(4)\prime}(r)$; for brevity we record the sum :

$$\begin{aligned}V^{(4)\prime}(r)+V^{(4)\prime\prime}(r) = {} & -\left(\frac{f^2}{4\pi}\right)^2\left(\frac{2\mu}{\pi x^3}\right)\\ & \times\{S_{12}[12\mathrm{K}_0(2x)+(15+4x^2)\,x^{-1}\,\mathrm{K}_1(2x)]\\ & -\boldsymbol{\sigma}^{(1)}.\boldsymbol{\sigma}^{(2)}\,[12\mathrm{K}_0(2x)+(12+8x^2)\,x^{-1}\,\mathrm{K}_1(2x)]\\ & +\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}\,[(23+4x^2)\,\mathrm{K}_0(2x)+(23+12x^2)\,x^{-1}\,\mathrm{K}_1(2x)]\}.\end{aligned} \qquad \ldots\ldots(67)$$

† In the Brillouin–Wigner perturbation series for the energy shift ΔE, this is a ΔE-dependence.

The expressions above are often taken as a first approximation to the one- and two-pion exchange potentials; use of the renormalized coupling constant partly accounts for radiative vertex corrections, but not corrections such as figure 14(*a*). All the methods described in § 3.2 agree here. For various derivations, see Taketani *et al.* (1952), Nishijima (1951), Klein (1953 b), Brueckner and Watson (1953 b), Feldman (1955) and Iwadare (1955).

There is less agreement about the choice of coupling, the use of a cut-off, and what terms to include. Several proposed potentials, and their agreement with experiment, are described in § 3.3.6.

With the coupling constants found necessary, the fourth-order contributions above are relatively strong compared to the second-order potential $V^{(2)}$, and prove important well beyond their characteristic range $r = (2\mu)^{-1}$. In fact, they cannot be ignored for $\mu r \lesssim 1{\cdot}5$. This raises doubts whether higher-order contributions converge at any useful radius; however, see § 3.3.7.

3.3.6. *Static potential models.*

The Taketani–Machida–Onuma (TMO) potential. These authors (Taketani *et al.* 1952) choose pseudovector coupling, and take for the outer potential $V^{(2)} + V^{(4)\prime} + V^{(4)\prime\prime}$ (also calculated by Nishijima 1951). They are able to fit the low-energy data with some freedom in the choice of parameters.

Fujii *et al.* (1954) study scattering up to 90 Mev, and find only qualitative agreement with experiment; they also correct some errors in the original paper of Taketani *et al.* (1952). The inner potential in triplet even states is a purely central attractive square well, of depth 125 Mev and radius $r_0 = 0{\cdot}6\ \mu^{-1}$. For all other states it is a hard core, with $r_0 = 0{\cdot}33\ \mu^{-1}$. The coupling constant is $f^2/4\pi = 0{\cdot}08$. There is fair agreement with n–p polarization at 100 Mev (Watari 1958; an earlier report to the contrary (Iwadare *et al.* 1956 a) is in error). The TMO potential is illustrated in figures 8–11.

The Lévy potential. Lévy (1952 b) calculates the leading adiabatic pseudoscalar terms, but wrongly finds an extra two-pair contribution

$$V_{2p}^{(4)\prime} = -12\mu\left(\frac{g^2}{4\pi}\right)^2\left(\frac{\mu}{2M}\right)^3\left[\frac{K_1(x)}{\pi x}\right]^2, \qquad \ldots\ldots(68)$$

and supposes the one-pair contribution to vanish. He takes the potential $V^{(2)} + V_{2p}^{(4)} + V_{2p}^{(4)\prime}$, with a hard core inner region, and represents some uncertain higher-order corrections by giving the second- and fourth-order potentials independent coupling constants g and G. The choice $g^2/4\pi = G^2/4\pi = 10{\cdot}36$, with core radius $r_0 = 0{\cdot}38\mu^{-1}$, gives a fair fit to data below 40 Mev (Lévy 1952 b, Martin and Verlet 1953), though the deuteron quadrupole moment is rather small. Later calculations show that a spin-independent core radius will not fit the low-energy data accurately (Preston and Shapiro 1957).

Klein (1953 a) and others point out Lévy's mistakes. Klein (1953 b) suggests the potential $V^{(2)} + V_{2p}^{(4)} + V_{1p}^{(4)}$, but finds it unsatisfactory (postscript to same paper).

The potential $V^{(2)} + V_{2p}^{(4)}$, again with two coupling constants, is also considered. It can be made to fit the low-energy data, but conflicts with high-energy experiments (Blatt and Kalos 1953, Jastrow 1953, Martin and Verlet 1954, Wertheim *et al.* 1956, Gelernter 1957).

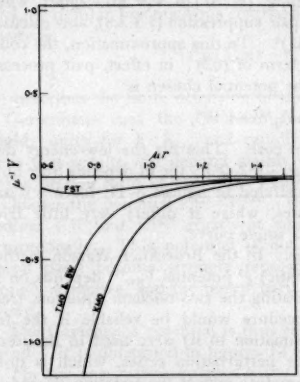


Figure 8. Static potential models in singlet even-parity states.

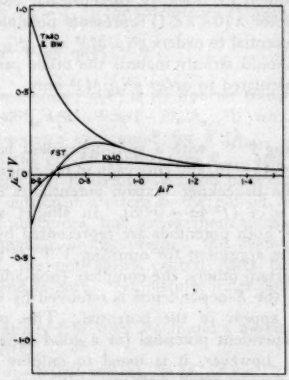


Figure 9. Static potential models in singlet odd-parity states.

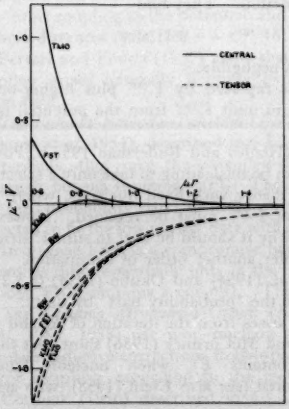


Figure 10. Static potential models in triplet even-parity states.

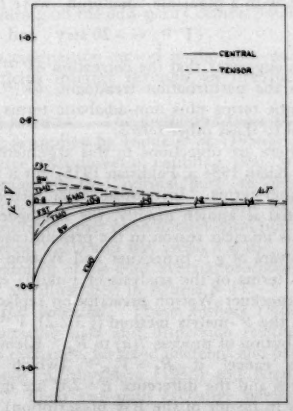


Figure 11. Static potential models in triplet odd-parity states.

Lévy and Marshak (1954) give a survey of these potentials.

The Brueckner–Watson (BW) potential. Brueckner and Watson (1953 b) use a modified fixed-source pseudoscalar coupling, $(g/2M)[\boldsymbol{\sigma}.\boldsymbol{\nabla}(\boldsymbol{\tau}.\boldsymbol{\varphi})+\lambda g\boldsymbol{\varphi}^3]$, where the factor λ $(0\leqslant\lambda\leqslant 1)$ represents possible pair suppression (§ 3.3.8), and calculate the potential to orders $g^2(\mu/M)^2$ and $g^4(\mu/M)^4$. To this approximation, the coupling should strictly include the other pair term of (62); in effect, pair processes are calculated to order $g^4(\mu/M)^3$ only. The potential chosen is

$$V^{(2)}+\lambda^2 V_{2p}{}^{(4)}+\lambda V_{1p}{}^{(4)}+V^{(4)\prime\prime},$$

omitting $V^{(4)\prime}$, with a spin-dependent hard core. This fits the low-energy data with $g^2/4\pi\sim 16$, and core radii in the range $\mu r_0\sim 0{\cdot}2$–$0{\cdot}3$; the fit depends little on λ.

The Brueckner–Watson potential is illustrated in figures 8–11, for $\lambda=0$ and $g^2/4\pi=14$ $(f^2/4\pi=0{\cdot}08)$. In singlet states, where it differs very little from TMO, both potentials are represented by a single curve.

The argument for omitting $V^{(4)\prime}$ is this. In the Brueckner–Watson method and certain others, the complete (non-adiabatic) g^2 potential $V_{BW}{}^{(2)}$ depends on E. When the E-dependence is removed by iterating the two-nucleon equation, fresh terms appear in the potential. This procedure would be reliable if the full E-independent potential (or a good approximation to it) were used in the iterations; however, it is usual to employ the perturbation series, which is quite misleading at small distances. We may therefore expect the iteration procedure to converge poorly, if at all. In the adiabatic limit, the leading correction term generated in this way is $V^{(4)\prime}$; we may expect it to be a poor approximation to the full correction.

Brueckner and Watson write $V_{BW}{}^{(2)}=V^{(2)}+\delta V^{(2)}$, and calculate expectation values with a plausible 'deuteron' wave function. They find

$$\langle V^{(2)}\rangle=-20\text{ Mev}\quad\text{and}\quad\langle\delta V^{(2)}\rangle=-0{\cdot}11\text{ Mev},$$

which suggests† that the correction $\delta V^{(2)}$ is negligible.

In the perturbation treatment, $\delta V^{(2)}$ is replaced by $V^{(4)\prime}$ plus higher-order adiabatic terms plus non-adiabatic terms; to omit $V^{(4)\prime}$ from the potential is to appeal to these other terms.

There are objections to this argument (Henley and Ruderman 1953 b, Power 1954, Klein 1954 a, Feldman 1955), for it can be misleading to take only a selection of higher terms. In neutral scalar meson theory with fixed nucleons, where the potential is known exactly, the analogue of $V^{(4)\prime}$ must be retained. However, there is no clear reason in the present case why it should be best to sum V strictly by powers of g; Brueckner and Watson offer another order of summation.

In terms of the analysis of Fukuda *et al.* (1954) and Okubo (1954) (§ 3.2.3), the Brueckner–Watson prescription replaces the 'probability part' by unity.

In the $\mathscr{S}$-matrix method (§ 3.2.2), $V^{(4)\prime}$ arises from the iteration of R_2 and the contribution of process 7(*a*) to R_4. Klein and McCormick (1956) show that these effects cancel if $R_2=V_{BW}{}^{(2)}$ (which contains E), when nucleon kinetic energies and the difference $E-2M$ are ignored (see also Klein (1958), who again argues in favour of the BW prescription).

† For a complete proof, higher-order perturbations and the effects on other deuteron properties and scattering would have to be studied.

The Gartenhaus potential. Gartenhaus (1955) takes fixed, extended nucleons with pseudovector coupling, encouraged by the success of this model in describing low-energy photoproduction and P-wave scattering of pions (Chew 1954, Chew and Low 1956). For nucleons centred at $\mathbf{r}_n$ ($n = 1, 2$), the interaction Hamiltonian density is

$$(f/\mu)\sum_n \rho(\mathbf{r}-\mathbf{r}_n)\boldsymbol{\sigma}^n.\boldsymbol{\nabla}(\boldsymbol{\tau}^n.\boldsymbol{\varphi}(\mathbf{r})). \qquad \ldots\ldots(69)$$

$\rho(\mathbf{r})$ describes the finite extension of the nucleons; $v(\mathbf{k})$ is its Fourier transform.

Gartenhaus uses the Gaussian function $v(\mathbf{k}) = \exp(-k^2/2k_{\max}^2)$, which is roughly unity for $k < k_{\max}$ and rapidly becomes very small for $k > k_{\max}$. (One hopes that results will depend mainly on the cut-off momentum $k_{\max}$, and little on the precise form of cut-off.) Following the Brueckner–Watson prescription, he chooses the potential $V_c^{(2)} + V_c^{(4)\prime}$; the subscript c refers to the cut-off, which modifies $V^{(2)}$ and $V^{(4)\prime}$ greatly at small distances, but leaves them unchanged asymptotically. This potential is not singular at the origin; the even-parity central parts become strongly repulsive, the odd-parity central parts become strongly attractive, and the tensor parts tend to zero. No distinct inner potential is used.

The Gartenhaus potential is thus essentially BW with $\lambda = 0$ and a momentum cut-off. It is illustrated in figures 12 and 13.

The low-energy data are fitted with $f^2/4\pi = 0{\cdot}089$, and $\omega_{\max} \equiv (k_{\max}^2 + \mu^2)^{1/2} = 6\mu$. These are close to the values required by Chew and Low (*loc. cit.*), which is gratifying. However, the strong attractive odd-parity core is wrong, leading to unphysical bound states, and the agreement with high-energy experiments is poor (Gammel and Thaler 1956, Signell and Marshak 1958).

Signell and Marshak (1957, 1958) achieve more success by adding an empirical spin–orbit coupling to the potential, and cutting off the odd-parity central potentials at short distances (§ 3.3.10).

Ferrari and Fonda (1957) show that a momentum cut-off in the pion–nucleon coupling arises naturally if these two fields interact through an intermediate boson field.

The Fukuda–Sawada–Taketani (FST) potential. The method of these authors (Fukuda *et al.* 1954), described in § 3.2.3, is applied by Inoue *et al.* (1956). Like Gartenhaus, they take the extended-source pseudovector coupling (69), with a Gaussian cut-off function $v(k)$. They calculate the 'normal part' to order g^4 and the 'probability part' to order g^2, with $f^2/4\pi = 0{\cdot}08$ and $k_{\max} = 6\mu$. This is expected to fit the low energy data with a suitable inner phenomenological potential.

In a sense FST is a compromise between TMO and BW (with $\lambda = 0$). These potentials differ essentially only in the treatment of the controversial processes 7(*a*) (though FST contains a cut-off). They are illustrated in figures 8–11.

The Konuma–Miyazawa–Otsuki (KMO) potential. These authors (Konuma *et al.* 1957, 1958) use pion–nucleon scattering data and dispersion relations to calculate radiative corrections to two-pion exchange, for fixed nucleons and pseudovector coupling, according to the methods of Miyazawa (1956) and Klein and McCormick (1956) described in § 3.2.2. Their two-pion exchange potential is $V^{(4)\prime} + V^{(4)\prime\prime}$ + radiative corrections (this follows Miyazawa rather than Klein and McCormick; the latter favour a Brueckner–Watson prescription which ignores

$V^{(4)\prime}$). Their one-pion exchange potential is $V^{(2)}$. They choose $f^2/4\pi = 0{\cdot}08$, and hard cores of radii $0{\cdot}49\mu^{-1}$ and $0{\cdot}5\mu^{-1}$ in triplet and singlet states, respectively. This potential agrees with the low-energy data and with p–p scattering at 18 Mev; including vacuum polarization effects (Eriksen *et al.* 1956) it gives a better fit to the low-energy triplet P-wave phase shifts than previous potentials (Otsuki and Tamagaki 1954; § 2.4.3).

The KMO potential is essentially TMO plus radiative corrections to two-pion exchange. Both are illustrated in figures 8–11.

General discussion. The number of proposed potentials (no one of which is outstandingly more plausible than the others) which differ in the two-pion exchange effects, the fact that two-pion exchange is important up to $\mu r = 1{\cdot}5$, and the fact that correction terms may be important in the same region (§ 3.3.7), suggest the following view.

For $\mu r > 1{\cdot}5$, the potential is close to the one-pion exchange potential $V^{(2)}$, with the renormalized coupling constant. For $0{\cdot}7 < \mu r < 1{\cdot}5$, the potential is *qualitatively* not unlike the suggested pion-theoretical potentials (which differ quite widely, but have many similarities). For $\mu r < 0{\cdot}7$, the potential is unknown, and must be treated empirically.

This view, less optimistic than the hopes expressed in § 3.3.2, is put forward by Iwadare *et al.* (1956 b, c, 1956). These authors, with Otsuki and Tamagaki (1955), show it to be consistent with the low-energy data, the low-energy triplet P-wave anomaly, p–p scattering at 18 Mev and (qualitatively) deuteron photo-disintegration at 20 Mev, taking $f^2/4\pi \sim 0{\cdot}08$. Otsuki (1958), Tamagaki (1958) and Watari (1958) go to higher energies and find that up to 150 Mev cross sections and polarizations also seem consistent with pion theory in this sense.

The TMO, BW, FST and KMO potentials are illustrated in figures 8 to 11, for the range $0{\cdot}6 < \mu r < 1{\cdot}5$. The Gartenhaus potential is illustrated separately in figures 12 and 13; the spin–orbit potential of Signell and Marshak (1957, 1958; § 3.3.10) is included in the latter figures.

3.3.7. *Correction terms.*

Many-pion exchange. Two-pion exchange effects are important as far as $r = 1{\cdot}5\mu^{-1}$, which led at first to the fear that many-pion exchange effects might not converge at any useful range. We expect greater convergence difficulties with pseudoscalar coupling, where the leading terms (in powers of μ/M) have an effective expansion parameter $(g^2/4\pi)(\mu/2M) \sim 1$. The parameter for pseudovector coupling is relatively small, $f^2/4\pi \sim 0{\cdot}1$.

In pseudoscalar coupling, Klein (1953 d) studies the leading adiabatic contributions of orders $(\mu/M)^{2n}$ and $(\mu/M)^{2n-1}$ from $2n$-pion exchange, and of order $(\mu/M)^{2n}$ from $(2n-1)$-pion exchange. All three series converge for $\mu r > X$, where $X\mathrm{e}^X = g^2\mu/4\pi M$. For $g^2/4\pi = 10$ or 15, $X = 0{\cdot}7$ or $0{\cdot}9$; with these values, convergence is slow for $\mu r < 1$.

In pseudovector coupling, Machida and Senba (1955) calculate the f^6 three-pion exchange effects, which prove small compared to one- and two-pion exchange effects for $\mu r > 0{\cdot}7$.

Radiative effects. By this we mean the effects of virtual pions which are not exchanged.

For pseudoscalar coupling, Lévy (1952 b) and Ruderman (1953) show that the leading corrections from vertex and self-energy parts are relatively small. However, a certain class of processes (e.g. figure 14(*a*)) has leading contributions (in

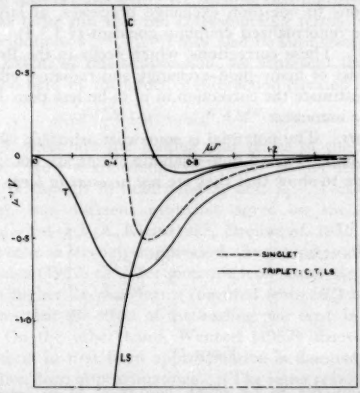


Figure 12. The Gartenhaus potential and the Signell–Marshak spin–orbit term in even-parity states.

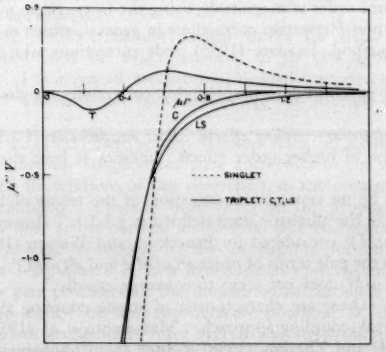


Figure 13. The Gartenhaus potential and the Signell–Marshak spin–orbit term in odd-parity states.

orders of μ/M) proportional to $V_{2p}{}^{(4)}$, and modifies this part of the potential by an unknown factor. Bonnevay (1954) shows that meson–meson scattering corrections (figure 14(*b*)) have a similar effect.

For pseudovector coupling, Henley and Ruderman (1953 b) estimate that radiative corrections to one- and two-pion exchange processes are relatively small. However, Konuma *et al.* (1958) calculate large corrections to two-pion exchange (compare the TMO and KMO potentials ; § 3.3.6 and figures 8–11).

Radiative corrections to one-pion exchange processes, at large distances, are included by using the renormalized coupling constant (§ 3.3.4).

Multiple scattering. These corrections, which occur in the Brueckner–Watson method (§ 3.2.5), consist of many-pion-exchange and radiative effects. Brueckner and Watson (1953 b) estimate the correction in v_1 to be less than 30% for $\mu r > 0{\cdot}6$, decreasing rapidly as r increases.

Non-adiabatic effects. The potential is somewhat arbitrary when we leave the adiabatic limit (§§ 3.2.6, 3.3.1), and non-adiabatic effects are not uniquely defined. However, one can hope to show that they are *not necessarily large* in some regions.

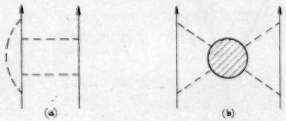


Figure 14.

Assuming nucleon momenta to be much less than M in intermediate states, I. Sato (1953) finds certain leading corrections for one- and two-pion exchange processes to be quite large at $\mu r = 1$ but negligible for $\mu r > 1{\cdot}5$. Eder (1954) finds corrections of the same order of magnitude, using the Lévy–Klein method (§ 3.2.4): however, this gives non-Hermitian corrections in general, which is unsatisfactory). With yet another method, Iwadare (1955) finds corrections which are negligible for $\mu r > 1$.

Because of their special interest, spin–orbit corrections are discussed in more detail later.

Specifically pseudoscalar-coupling effects. Pair suppression (§ 3.3.8), apparently due to a combination of higher-order effects, modifies at least the two-pair term $V_2{}^{(4)}$ greatly.

There seems to be no separate investigation of the terms of higher order in μ/M, which remain in the adiabatic limit defined in § 3.3.1. However, the no-pair terms of order $g^4(\mu/M)^4$ considered by Brueckner and Watson (1953 b ; § 3.3.6) are comparable with the pair terms of order $g^4(\mu/M)^2$ and $g^4(\mu/M)^3$; this particular series in powers of μ/M does not seem to converge rapidly.

Nuclear isobars. These are characteristic of strong-coupling theories, and do not appear in our weak-coupling approach. Matsumoto *et al.* (1953), Matsumoto and Sugawara (1954) and Zharkov (1958) include them phenomenologically, and find quite large effects.

Summary. These investigations are not exhaustive, but at least suggest that corrections cannot be ignored in the region $\mu r < 1{\cdot}5$, where two-pion exchange is important. However, they may not radically alter the qualitative features of the potential for $\mu r > 0{\cdot}7$.

3.3.8. *Pair suppression.*

In first Born approximation, the leading 'pair' term $g^2/2M\psi^* \gamma_4 \phi^2 \psi$ in the transformed pseudoscalar interaction (62) predicts much larger S-wave pion–nucleon scattering than is observed. We surmise that higher-order effects of some kind suppress this term and that the corresponding terms in nuclear forces (the leading pair contributions $V_{2p}^{(4)}, V_{1p}^{(4)}$) may also be suppressed.

To show the extent of this suppression, we represent the coupling of a fixed nucleon to the pion field by the effective interaction (treated in first Born approximation),

$$\rho_1(g^2/2M)\,\boldsymbol{\varphi}^2 + \rho_2(g^2/4M^2)\boldsymbol{\tau}.\boldsymbol{\varphi}\times\boldsymbol{\pi}. \qquad \ldots\ldots(70)$$

When $\rho_1 = \rho_2 = 1$, these are the leading pair terms of the transformed pseudoscalar interaction. Now a knowledge of S-wave pion–nucleon scattering in the low-energy limit is enough to determine ρ_1 and ρ_2. This scattering is not yet accurately known, but different estimates agree on the order of magnitude $\rho_1 = 0{\cdot}01$–$0{\cdot}03$, $\rho_2 = 0{\cdot}4$–$1{\cdot}3$ (A. Klein 1955, Drell *et al.* 1956, Okubo and Marshak 1958). The first term is strongly suppressed, the second is affected relatively little.

Drell and Henley (1952) consider pion–nucleon scattering, with fixed nucleons. Including certain higher $2n$-pion terms (omitted from (62) because of their order in M^{-1}), they show that the effect of the leading pair term is damped in first Born approximation. On the other hand, Wentzel (1952) observes that the effect of the leading pair term in first Born approximation is damped by the effects of the same term in higher Born approximations. (The same result is found by Edwards and Matthews (1957), who also exhibit damping in a relativistic treatment, including some radiative corrections.) These arguments appeal to different higher-order $\mathscr{S}$-matrix processes, in the original γ_5-theory. In each case the disagreement with experiment is reduced.†

Wentzel (1942, 1952) considers nuclear forces for static nucleons with ϕ^2 coupling only, and finds the exact potential to be $(1+Ag^2/4\pi M)^{-2}V_{2p}^{(4)}(r)$ at large distances. A is a cut-off momentum, needed for convergence, which is of the order of M. The two-pair contribution $V_{2p}^{(4)}$ is strongly suppressed, in this approximation.

Brueckner *et al.* (1953) observe that Wentzel takes no account of renormalization. They consider the simplest radiative correction to the nucleon propagator (figure 15(*a*)) and all iterations of this correction, in the renormalized theory with the original γ_5-coupling. There is little effect on the propagator for a nucleon with momentum near a free-particle momentum (the correction vanishes for free nucleons), but a large reduction in the propagator for an antinucleon with momentum near $-P_\mu$, where P_μ is a free-nucleon momentum. This correction therefore damps pair processes in the interaction of nucleons with pions of low momentum (and hence in the two-nucleon potential at distances $r \gg M^{-1}$). However, we cannot immediately infer overall pair suppression, for other corrections must be included. In threshold photo-pion production, this damping is completely cancelled by a correction at the photon–nucleon vertex (Kroll and Ruderman 1954, Deser *et al.* 1954). Again Hamada and Shono (1955) show that the correction 15(*b*), with all iterations, reduces the effect of 15(*a*).

† However, neither PS nor PV coupling (the latter gives S-wave scattering through recoil effects; Henley and Ruderman 1953 a) has yet given quantitative agreement here.

Klein (1954 c) considers the leading all-pair contribution (in powers of μ/M) to the adiabatic potential, including all radiative corrections to the nucleon propagators but neglecting pion–pion scattering and corrections to the pion propagators. He shows the two-pion exchange contribution to be simply $V_{2p}^{(4)}$ with g replaced

Figure 15.

by g_s, the special renormalized coupling constant of Deser *et al.* (1954), defined by the scattering of zero-energy pions by nucleons. $g_s^2/4\pi$ is found experimentally to be of order unity, which indicates a net pair suppression. (For a discussion of g and g_s, see Bethe and de Hoffmann 1955.)

Although no single argument is completely rigorous, there is a strong indication that at least the leading pair effects are suppressed in both pion–nucleon scattering and nuclear forces.

3.3.9. *Spin–orbit coupling.*

To fit high-energy scattering data, the two-nucleon potential seems to require something more than static central and tensor terms, such as an $\mathbf{L}.\mathbf{S}$ spin–orbit term (§§ 2.4.6, 2.6, 2.8, 3.3.10); nuclear structure suggests the same. It is interesting to see whether pion theory predicts a term of this kind.

We may expect $\mathbf{L}.\mathbf{S}$ terms on general grounds, if only as relativistic Thomas corrections. Klein (1953 b) exhibits such a correction to the two-pair term, $\mathbf{L}.\mathbf{S}(2M^2r)^{-1}\,\partial/\partial r\, V_{2p}^{(4)}(r)$; this has the wrong sign, however. Many other authors derive $\mathbf{L}.\mathbf{S}$ potentials, using a large variety of methods and approximations (Araki 1951, 1955, I. Sato 1953, Dresner 1953, Eder 1954, Shindo and Nishijima 1955, Iwadare 1955, S. Sato 1955, I. Sato *et al.* 1955, Okubo and Marshak 1958).† We simply record some leading terms in perturbation theory.

With pseudovector coupling, the leading $\mathbf{L}.\mathbf{S}$ correction is of relative order μ/M (compared with the usual $(\mu/M)^2$ Thomas correction). There is no f^2 contribution to this, but an f^4 contribution from two-pion exchange

$$-2\mu(\mu/M)(f^2/4\pi)^2(3+2\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)})(1+x)(1+x+x^2)x^{-6}\,e^{-2x}\,\mathbf{L}.\mathbf{S} \qquad \ldots\ldots(71)$$

(S. Sato 1955, I. Sato *et al.* 1955, Okubo and Marshak 1958). The controversial (§ 3.3.6) processes of figure 7(*a*) do not contribute in this order.

With pseudoscalar coupling, this term arises from the no-pair processes (replacing f by $g\mu/2M$). To the same order $g^4(\mu/M)^5$, there are also contributions from the pair processes, coming from the last two terms in (62). The ϕ^2 term gives the Thomas correction to $V_{2p}^{(4)}$ already recorded above; however, if we assume that pair suppression (§ 3.3.8) introduces a damping factor $\rho_1^2 \sim 10^{-3}$,

† There is also a qualitative discussion by Breit (1958 a), using an unquantized meson field.

this becomes negligible compared with (71). There remains a one-pair correction, coming from the (practically undamped) $\boldsymbol{\tau}.\boldsymbol{\varphi}\times\boldsymbol{\pi}$-term

$$-8\mu\left(\frac{g^2}{4\pi}\right)^2\left(\frac{\mu}{2M}\right)^5\boldsymbol{\tau}^{(1)}.\boldsymbol{\tau}^{(2)}\,(1+x)^2\,x^{-6}\,\mathrm{e}^{-2x}\,\mathbf{L}.\mathbf{S} \qquad \ldots\ldots(72)$$

(Okubo and Marshak 1958). All other g^4 corrections are of higher order in μ/M.

There is no reason to suppose that higher-order contributions converge rapidly, either in powers of g^2 or of μ/M. However, though they are too small, both (71) and (72) have roughly the order of magnitude required by the Gammel–Thaler and Signell–Marshak potentials (§§ 2.4.6, 2.8, 3.3.10); the signs are right for $T=1$ but wrong for $T=0$ states.

The theory also predicts other kinds of spin–orbit coupling, which are not usually considered. Okubo and Marshak (1958) derive quadratic $(\boldsymbol{\sigma}^{(1)}.\mathbf{L})(\boldsymbol{\sigma}^{(2)}.\mathbf{L})$ and $(\boldsymbol{\sigma}^{(1)}.\mathbf{P})(\boldsymbol{\sigma}^{(2)}.\mathbf{P})$ correction terms of relative order $(\mu/M)^2$ in both g^2 and g^4 potentials.

Okubo has shown how to calculate some effects of radiative corrections on the spin–orbit potential, using dispersion relations and an extension of the techniques of Miyazawa (1956) and Klein and McCormick (1956); see § 3.2.2 (private communication from S. Okubo).

3.3.10. *The Signell–Marshak potential.*

Signell and Marshak (1957, 1958) show that the static Gartenhaus potential (§ 3.3.6), which conflicts with high-energy scattering data, can be made to fit the latter quite well up to 150 Mev if an empirical spin–orbit potential $V_{LS}(r)\,\mathbf{L}.\mathbf{S}$ is added. They choose

$$V_{LS}(r) = V_0 y^{-1}\,\partial/\partial y[y^{-1}\exp(-y)], \quad (y = r/r_0), \qquad \ldots\ldots(73)$$

cut off at the constant value $V_{LS}(r_c)$ for $r<r_c = 0{\cdot}21\times10^{-13}$ cm, with $V_0 = 30$ Mev and $r_0 = 1{\cdot}07\times10^{-13}$ cm. This potential is taken from the work of Goldfarb and Feldman (1952) on the phenomenological model of Case and Pais (1950; § 2.4.4), except for the form of cut-off; it is simply a first estimate and has no deep significance.

Signell and Marshak also introduce a zero cut-off (i.e. put $V=0$) at short distances in the odd-parity Gartenhaus potential to remove some unphysical bound states which are otherwise predicted.

The possible effects of the spin–orbit force on the deuteron magnetic moment have caused some discussion. Feshbach (1957) estimates a large effect, which destroys the agreement secured by the Gartenhaus potential. As a remedy, de Swart *et al.* (1957) suggest that $V_{LS}(r)$ be made isobaric-spin dependent. If a weak long-range repulsive part is added in $T=0$ states, the effect in question can be removed without appreciably changing the fit to high-energy scattering; the spin–orbit term may even be taken to vanish altogether in these states without greatly damaging the fit to data. On the other hand, Soper (to be published) argues that the interaction moment from an $\mathbf{L}.\mathbf{S}$ potential is not well defined, and may be zero in the deuteron (§ 2.4.1).

The potential fits data up to 150 Mev quite well, though by no means perfectly (e.g. Fischer *et al.* (1958) criticize the fit to p–p scattering in singlet even states),

and also seems consistent with proton–carbon scattering data (Ohnuma 1958). A modified version, in which the $T = 0$ spin–orbit term vanishes and there is a finite repulsive core in odd states, has been used to explain deuteron photodisintegration in the 20–80 MeV range (de Swart and Marshak 1958, Zernik *et al.* 1959 ; § 2.4.5).

Later investigation shows that a smaller strength ($V_0 = 21$ MeV) and shorter range ($r_0 = 0{\cdot}7 \times 10^{-13}$ cm, as suggested by pion field theory § 3.3.9) for the spin–orbit term gives somewhat better agreement with experiment.† Unphysical bound states are now avoided by putting a hard core of radius $0{\cdot}37 \times 10^{-13}$ cm in the odd-parity potential. To distinguish it from the original Signell–Marshak (or SM) potential, this version is designated SM1 (Signell *et al.* 1958).

The Signell–Marshak potential is qualitatively similar to that proposed by Gammel and Thaler (1957 ; § 2.8) on purely phenomenological grounds, though there are many points of difference in detail. The Gartenhaus potential and the spin–orbit term (73) are shown in figures 12 and 13, the Gammel–Thaler potential in figures 1 and 2.

3.3.11. *Dispersion relations.*

The principle of causality implies certain general conditions, known as dispersion relations, on the predictions of a local field theory. In a simple example, the real part of the forward scattering amplitude at a given energy is related to the integral of the imaginary part of this amplitude, multiplied by certain factors, over all energies. Such relations have been applied, for example, to pion–nucleon scattering.

The dispersion relations for nucleon–nucleon scattering, discussed in detail by Goldberger, Nambu and Oehme (1957 ; see also Kuni 1956 and Grisaru 1958) are more difficult to apply. The general relativistic dispersion relations couple nucleon–nucleon to nucleon–antinucleon scattering. However, this coupling vanishes in a non-relativistic approximation, and if a further static approximation for the nucleons is introduced, the dispersion relations relate the Fourier transform of the two-nucleon potential to observable amplitudes. This approach to nuclear forces is urged by Matsuyama and Miyazawa (1958)‡ ; however, the nucleon–nucleon phase shifts which define the required amplitudes are not yet accurately known over the wide range of energies that is needed. (Other ways of deriving a potential from phase shifts are discussed in § 2.7.)

We have already noted in § 3.2.2 that pion–nucleon dispersion relations can be used to determine the adiabatic two-pion exchange potential from scattering data, assuming fixed nucleons with pseudovector coupling. This method is applied by Konuma *et al.* (1957, 1958 ; § 3.3.6), and extended to calculate $\mathbf{L}.\mathbf{S}$ spin–orbit corrections of relative order μ/M by Okubo (unpublished).

A more general type of dispersion relation, recently proposed by Mandelstam (1958 a, b), has interesting possibilities. This approach has already been used to deduce the pion–nucleon coupling constant directly from experiment (Chew

† This modified spin–orbit potential leads to roughly correct level splittings when inserted directly in the nuclear shell model, whereas the original form (73) gives values several times too large (private communication from J. P. Elliott). However, Nigam and Sundaresan (1958 b) using a different method, estimate that (73) gives too small a splitting.

‡ But see Hamilton (1959) for some difficulties with exchange potentials.

1958, Cziffra and Moravcsik 1958, Taylor *et al.* 1958) and to derive approximate results for nuclear–nucleon partial wave amplitudes (private communication from H. P. Noyes).

3.4. *Outlook*

The success of the semi-pion-theoretical potential of Signell and Marshak (1957 ; § 3.3.10) is real, but we may ask how much it is due to the phenomenological elements included in the potential.

When all is said, the predictions of pion theory itself are still somewhat limited. The asymptotic form of the potential, determined by one-pion exchange, is given unambiguously ; it seems consistent with experiment, but is relatively unimportant in most problems. On the other hand, at medium range ($r \sim \mu^{-1}$) the potential is not known at all well ; it seems unlikely that present methods will ever determine it reliably in this region. At short range ($\lesssim 0{\cdot}7\mu^{-1}$) we must resort to a wholly phenomenological description.

Of present techniques, those which relate whole sets of terms in the potential to observable quantities seem to be the most promising. They include the method of Miyazawa (1956) and Klein and McCormick (1956) (§ 3.2.2) for deducing the adiabatic two-pion exchange contribution from pion–nucleon scattering, and the application of nucleon–nucleon dispersion relations suggested by Matsuyama and Miyazawa (1958 ; § 3.3.11) ; unfortunately, both of these have to assume fixed nucleons.

The essentially non-relativistic concept of a potential, on which our present discussion is based, is both valid and useful at lower energies but becomes questionable at higher energies. For further progress it may be necessary to work instead directly with dispersion relations and other tools of relativistic field theory.

However, many physicists feel that the shortcomings of present field theory in general are due not only to inadequate mathematics but also to inadequate physics. A new physical theory of the elementary particles themselves may be needed before we gain a more detailed fundamental understanding of nuclear forces.

Acknowledgments

I would like to thank my colleagues at Harwell, and in particular Dr. J. S. Bell and Dr. T. H. R. Skyrme, for many valuable discussions during the preparation of this review. I am also grateful to a number of authors, mentioned in the text, for communicating unpublished results.

References

Abashian, A., and Hafner, E. M., 1958, *Phys. Rev. Let.*, **1**, 255.
Abraham, G., 1956, *Nuclear Phys.*, **1**, 415.
Anderson, H. L., Davidon, W. C., Glicksman, M., and Kruse, U. E., 1955, *Phys. Rev.*, **100**, 279.
Araki, G., 1951, *Progr. Theor. Phys., Japan*, **6**, 379 ; 1955, *Ibid.*, **13**, 13.
Arnowitt, R., and Deser, S., 1955, *Phys. Rev.*, **100**, 349.
Arnowitt, R., and Gasiorowicz, S., 1954, *Phys. Rev.*, **94**, 1057.
Ashkin, J., and Wu, T. Y., 1948, *Phys. Rev.*, **73**, 973.
Ashmore, A., Diddens, A. N., Huxtable, G. B., and Skarsvåg, K., 1958, *Proc. Phys. Soc.*, **72**, 289.

ASHMORE, A., DIDDENS, A. N., and HUXTABLE, G. B., 1959, *Proc. Phys. Soc.*, in the press.
AUSTERN, N., 1957, *Phys. Rev.*, **108**, 973.
BAKER, A. R., and WILKINSON, D. H., 1958, *Phil. Mag.*, **3**, 647.
BALDIN, A., 1958, *Nuovo Cim.*, **8**, 569.
BARGMANN, V., 1949 a, *Phys. Rev.*, **75**, 301 ; 1949 b, *Rev. Mod. Phys.*, **21**, 488.
BERETTA, L., CLEMENTEL, E., and VILLI, C., 1955 a, *Nuovo Cim.*, **1**, 739 (L) ; 1955 b, *Phys. Rev.*, **98**, 1526 (L).
BERGER, J. M., FOLDY, L. L., and OSBORN, R. K., 1952, *Phys. Rev.*, **87**, 1061.
BERNSTEIN, J., 1957, *Phys. Rev.*, **108**, 349.
BETHE, H. A., 1958, *Annals of Physics*, **3**, 190.
BETHE, H. A., and DE HOFFMANN, F., 1955, *Mesons and Fields*, Vol. 2 (New York : Row, Peterson).
BETHE, H. A., and LONGMIRE, C., 1950, *Phys. Rev.*, **77**, 647.
BETHE, H. A., and MORRISON, P., 1956, *Elementary Nuclear Theory*, 2nd Edn (New York : Wiley).
BIEDENHARN, L. C., and BLATT, J. M., 1954, *Phys. Rev.*, **93**, 1387.
BIEDENHARN, L. C., BLATT, J. M., and KALOS, M. H., 1958, *Nuclear Phys.*, **6**, 359.
BISHOP, G. R., and WILSON, R., 1957, *Handbuch der Physik*, Vol. 42 (Berlin : Springer).
BLANKENBECLER, R., 1958, *Phys. Rev.*, **111**, 1684.
BLATT, J. M., and BIEDENHARN, L. C., 1952 a, *Phys. Rev.*, **86**, 399 ; 1952 b, *Rev. Mod. Phys.*, **24**, 258.
BLATT, J. M., and KALOS, M. H., 1953, *Phys. Rev.*, **92**, 1563.
BLATT, J. M., and WEISSKOPF, V. F., 1952, *Theoretical Nuclear Physics* (New York : Wiley).
BODANSKY, D., ECCLES, S. F., FARWELL, G. W., RICKEY, M. E., and ROBISON, P. C., 1959, *Phys. Rev. Let.*, **2**, 101.
BONNEVAY, G., 1954, *C.R. Acad. Sci., Paris*, **238**, 1641.
BRANSDEN, B. H., SMITH, K., and TATE, C., 1958, *Proc. Roy. Soc.*, A, **247**, 73.
BREIT, G., 1955, *Phys. Rev.*, **99**, ; 1581 1957, *Ibid.*, **106**, 314 ; 1958 a, *Ibid.*, **111**, 652 ; 1958 b, *Phys. Rev. Let.*, **1**, 200.
BREIT, G., and BOURICIUS, W. G., 1949, *Phys. Rev.*, **75**, 1029.
BREIT, G., EHRMAN, J. B., and HULL, M. H., 1955, *Phys. Rev.*, **97**, 1051.
BREIT, G., and HULL, M. H., 1953, *Amer. J. Phys.*, **21**, 184 ; 1955, *Phys. Rev.*, **97**, 1047.
BROWN, G. E., 1957, *Proc. Phys. Soc.*, A, **70**, 361.
BROWN, G. E., and KANELLOPOULOS, T. V., 1957, *Proc. Phys. Soc.* A, **70**, 690, 703.
BRUECKNER, K. A., 1955, *Phys. Rev.*, **97**, 1353 ; 1958, *Proceedings of the* 1957 *Rehovoth Conference* (Amsterdam : North-Holland).
BRUECKNER, K. A., and GAMMEL, J. L., 1958, *Phys. Rev.*, **109**, 1023.
BRUECKNER, K. A., GELL-MANN, M., and GOLDBERGER, M., 1953, *Phys. Rev.*, **90**, 476.
BRUECKNER, K. A., and WATSON, K. M., 1953 a, *Phys. Rev.*, **90**, 699 ; 1953 b, *Ibid.*, **92**, 1023.
BURCHAM, W. E., 1955, *Progress in Nuclear Physics*, Vol. 4 (London : Pergamon).
CASE, K. M., and PAIS, A., 1950, *Phys. Rev.*, **80**, 203.
CASTILLEJO, L., and SINGH, L. S., 1959, *Nuovo Cim.*, **11**, 131.
CERN CONFERENCE, 1958 (CERN Scientific Information Service, Geneva).
CHADAN, K., 1955, *J. Phys. Radium*, **16**, 843 ; 1957, *C.R. Acad. Sci., Paris*, **245**, 1597 ; 1958 a, *Ibid.*, **246**, 1513 ; 1958 b, *Nuovo Cim.*, **10**, 892.
CHAMBERLAIN, O., SEGRÈ, E., TRIPP, R. D., WIEGAND, C., and YPSILANTIS, T., 1957, *Phys. Rev.*, **105**, 288.
CHEW, G. F., 1954, *Phys. Rev.*, **95**, 1669 ; 1958, *Ibid.*, **112**, 1380.
CHEW, G. F., and LOW, F. E., 1956, *Phys. Rev.*, **101**, 1570, 1579.
CHEW, G. F., and WICK, G. C., 1952, *Phys. Rev.*, **85**, 636.
CHISHOLM, J. S. R., and TOUSCHEK, B. F. X., 1953, *Phys. Rev.*, **90**, 763.
CHOU, K.-C., and SHIROKOV, M. I., 1958, *J. Exp. Theor. Phys.*, **34**, 1230 (see *Soviet Physics, J.E.T.P.*, **7**, 851).
CHRISTIAN, R. S., and HART, E. W., 1950, *Phys. Rev.*, **77**, 441.
CHRISTIAN, R. S., and NOYES, H. P., 1950, *Phys. Rev.*, **79**, 85.
CINI, M., 1953, *Nuovo Cim.*, **10**, 526, 614.

CLEMENTEL, E., and VILLI, C., 1955, *Nuovo Cim.*, **2**, 356 (L), 1165 ; 1957, *Ibid.*, **5**, 1166.
CLEMENTEL, E., VILLI, C., and JESS, L., 1957, *Nuovo Cim.*, **5**, 907.
CONDON, E. U., and SHORTLEY, G. H., 1935, *Theory of Atomic Spectra* (Cambridge : University Press).
COOPER, L. N., 1955, *Phys. Rev.*, **100**, 362.
CZIFFRA, P., MACGREGOR, M. H., MORAVCSIK, M. J., and STAPP, H. P., 1958, *Californian Report*, UCRL-8510.
CZIFFRA, P., and MORAVCSIK, M. J., 1958, *Bull. Amer. Phys. Soc.*, **3**, 404.
CZYŻ, W., and SAWICKI, J., 1957, *Nuovo Cim.*, **5**, 45 ; 1958 a, *Phys. Rev.*, **110**, 900 ; 1958 b, *Nuclear Phys.*, **8**, 621.
DALITZ, R. H., 1952, *Proc. Phys. Soc.* A, **65**, 175 ; 1954, *Phys. Rev.*, **95**, 799.
DANCOFF, S. M., 1950, *Phys. Rev.*, **78**, 382.
DESER, S., THIRRING, W. E., and GOLDBERGER, M. L., 1954, *Phys. Rev.*, **94**, 711.
DRELL, S. D., FRIEDMAN, M. H., and ZACHARIASEN, F., 1956, *Phys. Rev.*, **104**, 236.
DRELL, S. D., and HENLEY, E. M., 1952, *Phys. Rev.*, **88**, 1053.
DRELL, S. D., and HUANG, K., 1953, *Phys. Rev.*, **91**, 1527.
DRESNER, L., 1953, *Phys. Rev.*, **91**, 201 (L).
DURAND, L., 1957, *Phys. Rev.*, **108**, 1597.
DYSON, F. J., 1948, *Phys. Rev.*, **73**, 929 ; 1949, *Ibid.*, **75**, 486, 1736 ; 1953, *Ibid.*, **90**, 994 (L), **91**, 421 (L), 1543.
EBEL, M. E., and HULL, M. H., 1955, *Phys. Rev.*, **99**, 1596.
EDER, G., 1954, *Z. Naturforsch.*, **9a**, 565.
EDWARDS, S. F., and MATTHEWS, P. T., 1957, *Phil. Mag.*, **2**, 176, 467.
EISENBUD, L., and WIGNER, E. P., 1941, *Proc. Nat. Acad. Sci., Wash.*, **27**, 281.
ELLIOTT, J. P., and LANE, A. M., 1954, *Phys. Rev.*, **96**, 1160 ; 1957, *Handbuch der Physik*, Vol. 39 (Berlin : Springer).
ERIKSEN, E., FOLDY, L. L., and RARITA, W., 1956, *Phys. Rev.*, **103**, 781.
FELDMAN, D., 1953, *Phys. Rev.*, **89**, 1159 (L) ; 1955, *Ibid.*, **98**, 1456.
FERRARI, F., and FONDA, L., 1957, *Nuovo Cim.*, **6**, 1510.
FESHBACH, H., 1957, *Phys. Rev.*, **107**, 1626.
FESHBACH, H., and LOMON, E. L., 1956, *Phys. Rev.*, **102**, 891.
FESHBACH, H., and SCHWINGER, J., 1951, *Phys. Rev.*, **84**, 194.
FEYNMAN, R. P., 1949, *Phys. Rev.*, **76**, 749, 769.
FEYNMAN, R. P., and SPEISMAN, G., 1954, *Phys. Rev.*, **94**, 500 (L).
FISCHER, C. R., PYATT, K. D., HULL, M. H., and BREIT, G., 1958, *Bull. Amer. Phys. Soc.*, **3**, 183.
FOLDY, L. L., 1951, *Phys. Rev.*, **84**, 168 ; 1953, *Ibid.*, **92**, 178 ; 1958, *Rev. Mod. Phys.*, **30**, 471.
FOLDY, L. L., and ERIKSEN, E., 1954, *Phys. Rev.*, **95**, 1048 ; 1955, *Ibid.*, **98**, 775.
FOLDY, L. L., and WOUTHUYSEN, S. A., 1950, *Phys. Rev.*, **78**, 29.
FOWLER, J. L., and BROLLEY, J. E., 1956, *Rev. Mod. Phys.*, **28**, 103.
FRIED, B. D., 1954, *Phys. Rev.*, **95**, 851 (L).
FRÖBERG, C. E., 1947, *Phys. Rev.*, **72**, 519 (L).
FUJII, S., IWADARE, J., OTSUKI, S., TAKETANI, M., TANI, S., and WATARI, W., 1954, *Progr. Theor. Phys.*, **11**, 11.
FUJITA, J., and MIYAZAWA, H., 1957, *Progr. Theor. Phys., Japan.*, **17**, 360, 366.
FUKUDA, N., SAWADA, K., and TAKETANI, M., 1954, *Progr. Theor. Phys., Japan.*, **12**, 156.
FULTON, T., and NEWTON, R. G., 1956, *Nuovo Cim.*, **3**, 677.
GAMMEL, J. L., CHRISTIAN, R. S., and THALER, R. M., 1957, *Phys. Rev.*, **105**, 311.
GAMMEL, J. L., and THALER, R. M., 1956, *Phys. Rev.*, **103**, 1874 ; 1957, *Ibid.*, **107**, 291, 1337 ; 1959, *Elementary Particle and Cosmic Ray Physics*, in the press.
GARREN, A., 1953, *Phys. Rev.*, **92**, 213 (L), **93**, 1587 ; 1954, *Ibid.*, **96**, 1707 (L) ; 1955, USAEC report NYO-7102, *Thesis* (Carnegie Institute of Technology) ; 1956, *Phys. Rev.*, **101**, 419.
GARTENHAUS, S., 1955, *Phys. Rev.*, **100**, 900.
GELERNTER, H., 1957, *Phys. Rev.*, **105**, 1068.
GEL'FAND, I. M., and LEVITAN, B. M., 1951 a, *Doklady Akad. Nauk SSSR*, **77**, 557 ; 1951 b, *Izvest. Akad. Nauk SSSR*, **15**, 309.

GELL-MANN, M., and GOLDBERGER, M. L., 1953, *Phys. Rev.*, **91**, 398.
GELL-MANN, M., and LOW, F. E., 1951, *Phys. Rev.*, **84**, 350.
GELL-MANN, M., and WATSON, K. M., 1954, *Ann. Rev. Nucl. Sci.*, **4**, 219.
GLASSGOLD, A. E., 1959, *Proceedings of the* 1958 *Paris Conference on Nuclear Physics* (Paris : Dunod).
GOEBEL, C. J., KARPLUS, R., and RUDERMAN, M. A., 1955, *Phys. Rev.*, **100**, 240.
GOLDBERG, A., 1958, *Phys. Rev.*, **112**, 618.
GOLDBERGER, M. L., NAMBU, Y., and OEHME, R., 1957, *Annals of Physics*, **2**, 226.
GOLDFARB, L. J. B., and FELDMAN, D., 1952, *Phys. Rev.*, **88**, 1099.
GOURDIN, M., and MARTIN, A., 1957 a, *C.R. Acad. Sci., Paris*, **244**, 1153, 1329, 1469 ; 1957 b, *Nuovo Cim.*, **6**, 757 ; 1958, *Ibid.*, **8**, 699.
GREEN, H. S., 1956, *Nucl. Phys.*, **1**, 360.
GREEN, H. S., and BISWAS, S. N., 1957, *Progr. Theor. Phys., Japan*, **18**, 121.
GRISARU, M. T., 1958, *Phys. Rev.*, **111**, 1719.
HALL, H. H., and POWELL, J. L., 1953, *Phys. Rev.*, **90**, 912.
HAMADA, T., and SHONO, Y., 1955, *Progr. Theor. Phys., Japan*, **13**, 102 (L).
HAMILTON, J., 1959, *Phys. Rev.*, in the press.
HASEGAWA, H., 1955, *Progr. Theor. Phys., Japan*, **13**, 47.
HENLEY, E. M., and RUDERMAN, M. A., 1953 a, *Phys. Rev.*, **90**, 719 (L) ; 1953 b, *Ibid.*, **92**, 1036.
HESS, W. N., 1958, *Rev. Mod. Phys.*, **30**, 368.
HIIDA, K., IWADARE, J., and MACHIDA, S., 1956, *Progr. Theor. Phys., Japan*, **15**, 189.
HILLMAN, P., JOHANSSON, A., and TIBELL, G., 1958, *Phys. Rev.*, **110**, 1218 (L).
HOFSTADTER, R., 1957, *Ann. Rev. Nucl. Sci.*, **7**, 231.
HOFSTADTER, R., BUMILLER, F., and YEARIAN, M. R., 1958, *Rev. Mod. Phys.*, **30**, 482.
HOLMBERG, B., 1952, *Nuovo Cim.*, **9**, 597.
HSIEH, S. H., 1956 a, *Progr. Theor. Phys., Japan*, **16**, 68 (L) ; 1956 b, *Nuovo Cim.*, **4**, 138 (L) ; 1957, *Progr. Theor. Phys., Japan.*, **18**, 637 ; 1958, *Ibid.*, **19**, 453 (L).
HSIEH, S. H., and NAKAGAWA, M., 1956, *Progr. Theor. Phys., Japan*, **15**, 79 (L).
HULL, M. H., EHRMAN, J. B., HATCHER, R. D., and DURAND, L., 1956, *Phys. Rev.*, **103**, 1047.
HULTHÉN, L., and SUGAWARA, M., 1957, *Handbuch der Physik*, Vol. 39, p. 1 (Berlin : Springer).
HÜPER, R., 1957, *Z. Naturforsch.*, **12a**, 295.
HYLLERAAS, E. A., 1948, *Phys. Rev.*, **74**, 48.
IMAMURA, T., 1955, *Progr. Theor. Phys., Japan*, **13**, 183.
INOUE, K., MACHIDA, S., TAKETANI, M., and TOYODA, S., 1956, *Progr. Theor. Phys., Japan*, **15**, 122.
IWADARE, J., 1955, *Progr. Theor. Phys., Japan*, **13**, 189, 218 (L), **14**, 16.
IWADARE, J., OTSUKI, S., SANO, M., TAKAGI, S., and WATARI, W., 1956, *Progr. Theor. Phys., Japan*, **16**, 658.
IWADARE, J., OTSUKI, S., TAMAGAKI, R., and WATARI, W., 1956 a, *Nuovo Cim.*, **4**, 1204 ; 1956 b, *Progr. Theor. Phys., Japan*, **16**, 455, 472, 604 ; 1956 c, *Progr. Theor. Phys., Japan., Suppl.*, **3**, 32.
JACOBSOHN, B. A., 1953, *Phys. Rev.*, **89**, 881.
JANCOVICI, B., 1957, *Phys. Rev.*, **107**, 631 (L) ; 1958, *Nuovo Cim.*, **7**, 290.
JANKUS, V. Z., 1956, *Phys. Rev.*, **102**, 1586.
JASTROW, R., 1951, *Phys. Rev.*, **81**, 165 ; 1953, *Ibid.*, **91**, 749.
JOHNSON, M. H., and TELLER, E., 1955, *Phys. Rev.*, **98**, 783.
JOHNSTON, L. H., and YOUNG, D. E., 1958, *Bull. Amer. Phys. Soc.*, **3**, 50.
JONES, D. P., MURPHY, P. G., and O'NEILL, P. L., 1958, *Proc. Phys. Soc.*, **72**, 429.
JOST, R., and KOHN, W., 1952 a, *Phys. Rev.*, **87**, 977 ; 1952 b, *Ibid.*, **88**, 382 ; 1953, *K. Danske Vidensk. Selsk., Mat.-fys. Medd.*, **27**, No. 9.
JOST, R., and PAIS, A., 1951, *Phys. Rev.*, **82**, 840.
KARPLUS, R., and WATSON, K. M., 1957, *Amer. J. Phys.*, **25**, 641.
KAWAGUCHI, M., 1958, *Phys. Rev.*, **111**, 1314.
KAY, I., and MOSES, H. E., 1955, *Nuovo Cim.*, **2**, 917 ; 1956, *Ibid.*, **3**, 66, 276 ; 1957, *Ibid.*, **5**, Suppl., 230.

KELLER, A., 1955, *Proc. Phys. Soc.* A, **68**, 930.
KEMMER, N., 1938, *Proc. Camb. Phil. Soc.*, **34**, 354.
KLEIN, A., 1953 a, *Phys. Rev.*, **89**, 1158 (L); 1953 b, *Ibid.*, **90**, 1101; 1953 c, *Ibid*, **91**, 1285 (L); 1953 d, *Ibid.*, **92**, 1017; 1954 a, *Ibid.*, **94**, 195; 1954 b, *Ibid.*, **94**, 1052; 1954 c, *Ibid.*, **95**, 1061; 1954 d, *Ibid.*, **95**, 1676; 1955, *Ibid.*, **99**, 998.
KLEIN, A., 1958, *Progr. Theor. Phys., Japan*, **20**, 257.
KLEIN, A., and MCCORMICK, B. H., 1956, *Phys. Rev.*, **104**, 1747; 1958, *Progr. Theor. Phys., Japan*, **20**, 876.
KLEIN, C. A., 1955, *Nuovo Cim.*, **1**, 581, **2**, 38.
KOHN, W., 1954, *Rev. Mod. Phys.*, **26**, 292, 472.
KONUMA, M., MIYAZAWA, H., and OTSUKI, S., 1957, *Phys. Rev.*, **107**, 320 (L); 1958, *Progr. Theor. Phys., Japan*, **19**, 17.
KRÓLIKOWSKI, W., and RZEWUSKI, J., 1955, *Nuovo Cim.*, **2**, 203; 1956, *Ibid.*, **3**, 260.
KROLL, N. M., and RUDERMAN, M. A., 1954, *Phys. Rev.*, **93**, 233.
KUNI, F. M., 1956, *Doklady Akad. Nauk S.S.S.R.*, **111**, 571 (see *Soviet Physics 'Doklady' (New York)*, **1**, 686).
KURŞUNOĞLU, B., 1954, *Phys. Rev.*, **96**, 1690.
LAING, E. W., 1955, *Phil. Mag.*, **46**, 106.
LAURITSEN, T., 1952, *Ann. Rev. Nucl. Sci.*, **1**, 67.
LEHMANN, H., 1953, *Z. Naturforsch.*, **8a**, 579.
LEPORE, J. V., 1952, *Phys. Rev.*, **88**, 750.
LEVINSON, N., 1949, *K. Danske Vidensk. Selsk., Mat.-fys. Medd.*, **25**, No. 9; 1953, *Phys. Rev.*, **89**, 755.
LÉVY, M. M., 1952 a, *Phys. Rev.*, **88**, 72; 1952 b, *Ibid.*, **88**, 725.
LÉVY, M. M., and MARSHAK, R. E., 1954, *Proceedings of the Glasgow Conference on Nuclear and Meson Physics* (London: Pergamon Press).
LIPPMANN, B. A., and SCHWINGER, J., 1950, *Phys. Rev.*, **79**, 469.
LOMON, E., 1958, *Bull. Amer. Phys. Soc.*, **3**, 183.
LOW, F. E., 1955, *Phys. Rev.*, **97**, 1392.
MCCAULEY, G. P., and BROWN, G. E., 1958, *Proc. Phys. Soc.*, **71**, 893.
MACGREGOR, M. H., 1958, *University of California Report UCRL-5263.*
MACHIDA, S., 1953, *Progr. Theor. Phys., Japan*, **9**, 683.
MACHIDA, S., and SENBA, K., 1955, *Progr. Theor. Phys., Japan*, **13**, 389.
MCINTYRE, J. A., 1956, *Phys. Rev.*, **103**, 1464.
MCINTYRE, J. A., and BURLESON, G. R., 1958, *Phys. Rev.*, **112**, 2077.
MCINTYRE, J. A., and DHAR, S., 1957, *Phys. Rev.*, **106**, 1074.
MACKE, W., 1953 a, *Phys. Rev.*, **91**, 195 (L); 1953 b, *Nuovo Cim.*, **10**, 1198 (L); 1953 c, *Z. Naturforsch.*, **8a**, 599, 615.
MCMANUS, H., and THALER, R. M., 1958, *Phys. Rev.*, **110**, 590 (L).
MCMANUS, H., THALER, R. M., and KERMAN, A. K., 1958, *Proceedings of the* 1958 *Conference on High-energy Physics at CERN, Session* 2.
MANDELSTAM, S., 1958 a, *Proceedings of the* 1958 *Conference on High-energy Physics at CERN, Session* 3; 1958 b, *Phys. Rev.*, **112**, 1344.
MANDL, F., 1957, *Quantum Mechanics* (London: Butterworths Scientific Publications); 1958, *Proc. Phys. Soc.*, **71**, 686.
MARTIN, A., and VERLET, L., 1953, *Phys. Rev.*, **89**, 519 (L); 1954, *Nuovo Cim.*, **12**, 483.
MARTIN, P. C., and DE DOMINICIS, C., 1957, *Phys. Rev.*, **105**, 1419 (L).
MASSEY, H. S. W., 1953, *Progress in Nuclear Physics*, Vol. 3 (London: Pergamon Press).
MATHER, K. B., and SWAN, P., 1958, *Nuclear Scattering* (Cambridge: University Press).
MATSUMOTO, T., HAMADA, T., and SUGAWARA, M., 1953, *Progr. Theor. Phys., Japan.*, **10**, 199.
MATSUMOTO, T., and SUGAWARA, M., 1954, *Progr. Theor. Phys., Japan*, **12**, 553.
MATSUYAMA, S., and MIYAZAWA, H., 1958, *Progr. Theor. Phys., Japan*, **19**, 517.
MATTHEWS, P. T., and SALAM, A., 1951, *Rev. Mod. Phys.*, **23**, 311.
MIYAZAWA, H., 1956, *Phys. Rev.*, **104**, 1741.
MØLLER, C., and ROSENFELD, L., 1939, *Nature, Lond.*, **143**, 241; 1940, *K. Danske Vidensk. Selsk., Mat.-fys. Medd.*, **17**, No. 8.
MORAVCSIK, M. J., 1958, *University of California Report*, UCRL-5317-T.

MORITA, M., SUGIE, A., and YOSHIDA, S., 1954, *Progr. Theor. Phys., Japan*, **12**, 713.
MOSZKOWSKI, S. A., and SCOTT, B. L., 1958, *Phys. Rev. Let.*, **1**, 298.
MOTT, N. F., and MASSEY, H. S. W., 1949, *Theory of Atomic Collisions*, 2nd Edn (Oxford : Clarendon Press).
NAKABAYASI, K., and SATO, I., 1952, *Phys. Rev.*, **88**, 144 (L).
NAMBU, Y., 1948, *Progr. Theor. Phys., Japan*, **3**, 444 ; 1950, *Ibid.*, **5**, 614.
NEWTON, R. G., 1955, *Phys. Rev.*, **100**, 412 ; 1956, *Ibid.*, **101**, 1588 ; 1957 a, *Ibid.*, **105**, 763 (L) ; 1957 b, *Ibid.*, **107**, 1025.
NEWTON, R. G., and FULTON, T., 1957, *Phys. Rev.*, **107**, 1103.
NEWTON, R. G., and JOST, R., 1955, *Nuovo Cim.*, **1**, 590.
NEWTON, R. G., and SCOFIELD, J. H., 1958, *Phys. Rev.*, **110**, 785 (L).
NICHOLSON, A. F., and BROWN, G. E., 1959, *Proc. Phys. Soc.*, **73**, 221.
NIGAM, B. P., and SUNDARESAN, M. K., 1958 a, *Canad. J. Phys.*, **36**, 571 ; 1958 b, *Phys. Rev.*, **111**, 284.
NISHIJIMA, K., 1951, *Progr. Theor. Phys., Japan*, **6**, 815, 911.
NOGAMI, Y., and HASEGAWA, H., 1956, *Progr. Theor. Phys., Japan.*, **15**, 137.
NOVOZHILOV, IU. V., 1957, *J. Exp. Theor. Phys.*, **33**, 901 (see *Soviet Physics, J.E.T.P.* (*New York*), **33** (6), 692 (1958)).
NOYES, H. P., and MACGREGOR, M. H., 1958, *Phys. Rev.*, **111**, 223.
OEHME, R., 1955, *Phys. Rev.*, **98**, 147, 216 (L).
OHMURA, T., 1956, *Progr. Theor. Phys., Japan*, **16**, 231.
OHNUMA, S., 1957, *Phys. Rev.*, **108**, 460 ; 1958, *Ibid.*, **111**, 1173.
OHNUMA, S., and FELDMAN, D., 1956, *Phys. Rev.*, **102**, 1641.
OKUBO, S., 1954, *Progr. Theor. Phys., Japan*, **12**, 102 (L), 603.
OKUBO, S., and MARSHAK, R. E., 1958, *Annals of Physics*, **4**, 166.
OTSUKI, S., 1958, *Progr. Theor. Phys., Japan*, **20**, 171.
OTSUKI, S., and TAMAGAKI, R., 1954, *Progr. Theor. Phys., Japan*, **12**, 806 ; 1955, *Ibid.*, **14**, 52.
PAULI, W., 1946, *Meson Theory of Nuclear Forces* (New York : Interscience).
PHILLIPS, R. J. N., 1957 a, *Nuovo Cim.*, **5**, 1335 ; 1957 b, *Proc. Phys. Soc.* A, **70**, 721 ; 1958, *Nuovo Cim.*, **8**, 265.
POWER, E. A., 1954, *Nuovo Cim.*, **12**, 323.
PRESTON, M. A., and SHAPIRO, J., 1954, *Phys. Rev.*, **96**, 813 (L) ; 1957, *Canad. J. Phys.*, **35**, 451.
PUZIKOV, L., RYNDIN, R., and SMORODINSKY, J., 1957, *Nuclear Phys.*, **3**, 436.
RAPHAEL, R. B., 1956, *Phys. Rev.*, **102**, 905 ; 1957, *Ibid.*, **107**, 1135.
REGGE, T., 1958, *Nuovo Cim.*, **9**, 491.
RIAZUDDIN, 1958, *Nuclear Phys.*, **7**, 217, 223.
RIESENFELD, W. B., and WATSON, K. M., 1956, *Phys. Rev.*, **102**, 1157.
ROCHESTER CONFERENCES 1955–6–7 (Interscience Publications, New York).
ROSEN, L., and BROLLEY, J. E., 1959, *Phys. Rev. Let.*, **2**, 98.
ROSENFELD, L., 1948, *Nuclear Forces* (Amsterdam : North-Holland).
RUDERMAN, M. A., 1953, *Phys. Rev.*, **90**, 183.
SACHS, R. G., 1953, *Nuclear Theory* (New York : Addison-Wesley).
SALPETER, E. E., 1952, *Phys. Rev.*, **87**, 328 ; 1953, *Ibid.*, **91**, 994.
SALPETER, E. E., and BETHE, H. A., 1951, *Phys. Rev.*, **84**, 1232.
SAPERSTEIN, A. M., and DURAND, L., 1956, *Phys. Rev.*, **104**, 1102.
SATO, I., 1953, *Progr. Theor. Phys.*, **10**, 323.
SATO, I., ITABASHI, K., and SATO, S., 1955, *Progr. Theor. Phys., Japan*, **14**, 303.
SATO, S., 1955, *Progr. Theor. Phys., Japan*, **13**, 457 (L).
SAWICKI, J., and WATANABE, S., 1959, *Nucl. Phys.*, **10**, 151.
SCHIFF, L. I., 1949, *Quantum Mechanics* (New York : McGraw-Hill).
SCHWEBER, S. S., BETHE, H. A., and DE HOFFMANN, F., 1955, *Mesons and Fields*, Vol. 1 (New York : Row, Peterson).
SCHWINGER, J., 1948, *Phys. Rev.*, **74**, 1439 ; 1950, *Ibid.*, **78**, 135 ; 1951, *Proc. Nat. Acad. Sci., Wash.*, **37**, 452, 455.
SESSLER, A. M., and FOLEY, H. M., 1958, *Phys. Rev.*, **110**, 995 (L).
SHAPIRO, J., and PRESTON, M. A., 1956, *Canad. J. Phys.*, **34**, 451.

SHARP, R. T., 1958, *Nuovo Cim.*, **9**, 23.
SHINDO, M., and NISHIJIMA, K., 1955, *Progr. Theor. Phys., Japan*, **13**, 103 (L).
SIEGEL, R. T., HARTZLER, A. J., and LOVE, W. A., 1956, *Phys. Rev.*, **101**, 838.
SIGNELL, P. S., and MARSHAK, R. E., 1957, *Phys. Rev.*, **106**, 832 (L); 1958, *Ibid.*, **109**, 1229.
SIGNELL, P. S., ZINN, R., and MARSHAK, R. E., 1958, *Phys. Rev. Let.*, **1**, 416.
SIMON, A., 1953, *Phys. Rev.*, **92**, 1050.
SKYRME, T. H. R., 1956, *Phil. Mag.*, **1**, 1043; 1959, *Nucl. Phys.*, **9**, 615, 635.
SOPER, J. M., 1957, *Phil. Mag.*, **2**, 1219.
STAFFORD, G. H., WHITEHEAD, C., and HILLMAN, P., 1957, *Nuovo Cim.*, **5**, 1589.
STAPP, H. P., 1955, *Thesis*, University of California, UCRL-3098; 1956, *Phys. Rev.*, **103**, 425.
STAPP, H. P., YPSILANTIS, T., and METROPOLIS, N., 1957, *Phys. Rev.*, **105**, 302.
STECH, D., 1956, *Z. Phys.*, **144**, 214.
STERNHEIMER, R. M., 1955, *Phys. Rev.*, **100**, 886.
STÜCKELBERG, E. C. G., 1938, *Phys. Rev.*, **54**, 889.
SUGAWARA, M., 1955 a, *Phys. Rev.*, **99**, 1601; 1955 b, *Ark. Fys.*, **10**, 113; 1955 c, *Progr. Theor. Phys., Japan*, **14**, 535.
SUGIE, A., 1954, *Progr. Theor. Phys., Japan*, **11**, 333 (L).
SUNDARESAN, M. K., 1957, *Phys. Rev.*, **105**, 1075.
SWANSON, D. R., 1953, *Phys. Rev.*, **89**, 740.
DE SWART, J. J., CZYŻ, W., and SAWICKI, J., 1959, *Phys. Rev. Let.*, **2**, 51.
DE SWART, J. J., and MARSHAK, R. E., 1958, *Phys. Rev.*, **111**, 272.
DE SWART, J. J., MARSHAK, R. E., and SIGNELL, P. S., 1957, *Nuovo Cim.*, **6**, 1189.
SYMANZIK, K., 1954, *Nuovo Cim.*, **11**, 88.
TAKEDA, G., and WATSON, K. M., 1954, *Phys. Rev.*, **94**, 1087 (L); 1955, *Ibid.*, **97**, 1336.
TAKETANI, M., MACHIDA, S., and ONUMA, S., 1952, *Progr. Theor. Phys., Japan*, **7**, 45.
TAKETANI, M., NAKAMURA, S., and SASAKI, M., 1951, *Progr. Theor. Phys., Japan*, **6**, 581.
TAMAGAKI, R., 1958, *Progr. Theor. Phys., Japan*, **20**, 505.
TAMM, I., 1945, *J. Phys., U.S.S.R.*, **9**, 449.
TANI, S., 1951, *Progr. Theor. Phys., Japan*, **6**, 267; 1954, *Ibid.*, **12**, 104.
TANNER, N., 1957, *Phys. Rev.*, **107**, 1203.
TAYLOR, A. E., 1957, *Rep. Progr. Phys.*, **20**, 86 (London: Physical Society).
TAYLOR, A. E., WOOD, E., and BIRD, L., 1958, *Proceedings of* 1958 *Conference on High-energy Physics at CERN, Session* 2.
TAYLOR, J. C., 1954 a, *Phys. Rev.*, **95**, 1313; 1954 b, *Ibid.*, **96**, 1438.
TAYLOR, J. G., MORAVCSIK, M. J., and URETSKY, J. L., 1958, *University of California Report*, UCRL-8317.
THALER, R. M., and BENGSTON, J., 1954, *Phys. Rev.*, **94**, 679.
THALER, R. M., BENSTON, J., and BREIT, G., 1954, *Phys. Rev.*, **94**, 683.
TOMONAGA, S., 1946, *Progr. Theor. Phys., Japan*, **1**, 27.
TRIPP, R. D., 1956, *Phys. Rev.*, **102**, 862.
VAIDYA, S. M., 1958, *Nucl. Phys.*, **5**, 179.
VAN HOVE, L., 1949, *Phys. Rev.*, **75**, 1519.
VERDE, M., 1957, *Nuovo Cim.*, **6**, 340.
WATARI, W., 1958, *Progr. Theor. Phys., Japan*, **20**, 181.
WATSON, G. N., 1944, *Theory of Bessel Functions*, 2nd Edn (Cambridge: University Press).
WATSON, K. M., and LEPORE, J. V., 1949, *Phys. Rev.*, **76**, 1157.
WATTENBERG, A., 1957, *Handbuch der Physik*, Vol. 40 (Berlin: Springer).
WENTZEL, G., 1942, *Helv. Phys. Acta*, **15**, 111; 1952, *Phys. Rev.*, **86**, 802 (L).
WERLE, J., 1955, *Nuovo Cim.*, **1**, 537.
WERTHEIM, M. S., HULL, M. H., and SAPERSTEIN, A. M., 1956, *Phys. Rev.*, **104**, 764.
WICK, G. C., 1938, *Nature, Lond.*, **142**, 993; 1955, *Rev. Mod. Phys.*, **27**, 339.
WIGNER, E. P., 1955, *Phys. Rev.*, **98**, 145.
WILKINSON, D. H., 1956, *Phil. Mag.*, **1**, 1031; 1958, *Phys. Rev.*, **109**, 1603, 1610, 1614.
WILSON, R., 1957, *Proceedings of the 7th Rochester Conference* (New York: Interscience).
WILSON, R. R., 1956, *Phys. Rev.*, **104**, 218.
DE WIT, M., and DURAND, L., 1958, *Phys. Rev.*, **111**, 1507.

WOLFENSTEIN, L., 1949, *Phys. Rev.*, **75**, 1664, **76**, 541 ; 1954, *Ibid.*, **96**, 1654 ; 1956 a, *Ibid.*, **101**, 427 ; 1956 b, *Ann. Rev. Nucl. Sci.*, **6**, 43 ; 1956 c, *Bull. Amer. Phys. Soc.*, **1**, 36.
WOLFENSTEIN, L., and ASHKIN, J., 1952, *Phys. Rev.*, **85**, 947.
WRIGHT, S. C., 1955, *Phys. Rev.*, **99**, 996.
YAMAGUCHI, Y., 1954, *Phys. Rev.*, **95**, 1628 ; 1958, *Progr. Theor. Phys., Japan*, **19**, 622.
YAMAGUCHI, Y., and YAMAGUCHI, Y., 1954, *Phys. Rev.*, **95**, 1635.
YENNIE, D. R., LÉVY, M. M., and RAVENHALL, D. G., 1957, *Rev. Mod. Phys.*, **29**, 144.
YNTEMA, J. L., and WHITE, M. G., 1954, *Phys. Rev.*, **95**, 1226.
YOVITS, M. C., SMITH, R. L., HULL, M. H., BENGSTON, J., and BREIT, G., 1952, *Phys. Rev.*, **85**, 540.
ZASTAVENKO, L. G., RYNDIN, R. M., and CHOU, G. C., 1958, *J. Exp. Theor. Phys.*, **34**, 526 (see *Soviet Physics, J.E.T.P.*, **7**, 363).
ZERNIK, W., RUSTGI, M. L., and BREIT, G., 1959, *Phys. Rev.*, in the press.
ZHARKOV, G. F., 1958, *J. Exp. Theor. Phys.*, **34**, 1211 (see *Soviet Physics, J.E.T.P.*, **7**, 837).
ZIMMERMAN, W., 1954, *Nuovo Cim.*, **11**, Suppl. **1**, 43.